Final Review

Ricky Truong

Final

- **Written Component**: Thur, 05/15 from 2 to 5 PM
- **Oral Component**: Over Zoom BEFORE (10 minute sessions)
- **You all got this!** 🙂

Logistics and Disclaimer

- This will be half lecture/content review (by Ricky and Tino) and half hands-on practice (by Sarah and Maggie)
- We couldn't fit every detail from the into 2 hours, so these are the main/important ideas!
- We don't know what the exam looks like!

Pre-Midterm Material

- We'll touch on some concepts pre-midterm, but the focus of this will be **post-midterm**
  - Slides from Midterm Review: https://drive.google.com/file/d/1Ro-Jw30YOg5et9TGaCWhLqteb5AaJ7S6/view?usp=drive_link

Before we start, what topics do you want me to spend the most time covering?

Content Review: Week 9

Foundations of Probability

- **Probability**: A value between 0 and 1 (intuitively, a "long-term frequency")
  - **Naive probability** is all favorable outcomes / all possible outcomes
  - *Ex: Probability of getting dealt an ace is 4/52 = 0.077*
- **Outcome**: Result after conducting an **experiment**
  - *Ex: After the experiment, I get dealt the ace of hearts*
- **Sample space**: Set of all possible **outcomes** of **experiment**
  - *Ex: There are 52 cards I could've been dealt*
- **Event**: Collection of **outcomes**
  - *Ex: The event I get dealt an ace is the collection of 4 specific outcomes*
  - *If A = the event I get dealt an ace, then P(A) = 0.077*

Recapping Our Toolkit: Notes

- **Union**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - For **disjoint** events, $P(A \cup B) = P(A) + P(B)$ because $P(A \cap B) = 0$
- **Intersection**: $P(A \cap B) = P(A)\,P(B \mid A) = P(B)\,P(A \mid B)$
  - For **independent** events, $P(A \cap B) = P(A)\,P(B)$ because $P(A \mid B) = P(A)$
- **Complement Rule**: $P(A) = 1 - P(A^C)$, $P(A \mid B) = 1 - P(A^C \mid B)$
  - Use when you see "**at least**" (e.g., "Find the probability of rolling a 5+ at least once in 3 rolls")
- **Def. of Conditional Probability**: $P(A \mid B) = P(A \cap B) / P(B)$
- **Bayes' Rule**: $P(A \mid B) = P(B \mid A)\,P(A) / P(B)$
- **LOTP**: $P(A) = P(A \mid B)\,P(B) + P(A \mid B^C)\,P(B^C)$
  - Use for **wishful thinking** (e.g., "I really wish I knew which factory the cone came from")
- In general with probability, **start by defining events**

One More Probability!

- **Positive predictive value (PPV)**: In a diagnostic test, the probability that a person has the disease, given that they tested positive for it (**true positive**)
- PPV = $P(D \mid T^+)$, where D is event of having disease and T^+ is event of testing positive

$$P(D|T^+) = \frac{P(D \cap T^+)}{P(T^+)} = \frac{P(D \cap T^+)}{P(D \cap T^+) + P(D^c \cap T^+)} = \frac{P(T^+|D)P(D)}{P(T^+|D)P(D) + P(T^+|D^c)P(D^c)}$$
$$= \frac{(\text{sens})(\text{prev})}{(\text{sens})(\text{prev}) + (1 - \text{spec})(1 - \text{prev})}$$

What are the 2 main tools for finding unconditional probability, such as P(A)?

Question:

What are the 2 main tools for finding unconditional probability, such as P(A)?

Complement rule and **LOTP**.

We often use complement for “at least” (e.g., “Find the probability of rolling a 5+ at least once in 3 rolls”).

We often use LOTP for “wishful thinking” (e.g., “I really wish I knew which factory the cone came from”).

Every upperclassman has a probability p of buying a scooter. If they live in the Quad, p = 1/10. Otherwise, p = 1/20. I want to find the probability a randomly selected upperclassman buys a scooter. What tool should I use?

Question:

Every upperclassman has a probability p of buying a scooter. If they live in the Quad, p = 1/10. Otherwise, p = 1/20. I want to find the probability a randomly selected upperclassman buys a scooter. What tool should I use?

LOTP (since "I wish" I knew whether or not they're in the Quad)! Let B be the event they buy a scooter and Q be the event they're in the Quad.

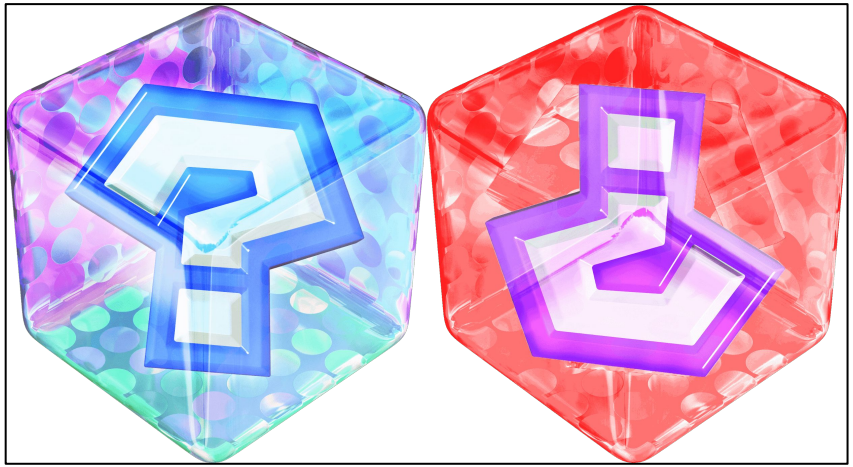
$P(B) = P(B \mid Q)\,P(Q) + P(B \mid Q^C)\,P(Q^C)$ by **LOTP**. $P(B) = (1/10)(3/12) + (1/20)(9/12) = 0.0625$.

Random Variables

- **Random variable**: A function that maps each **event** in the **sample space** to a number
- Intuitively, think of a r.v. as an unknown value that "crystallizes" to a certain number AFTER an **experiment**
  - *Ex: X is a r.v. for the number of heads I get after flipping 10 coins. X could be 0, 1, ..., or 10. After the experiment, it "crystallizes" to one of those numbers.*

A Silly (but Helpful) Intuition for Random Variables

- Think of **random variables** as mystery boxes in Mario Kart
- It's unknown what it will crystallize to, but we can still describe the random variable with probabilities
  - For example, there's a pretty low probability this random variable will crystallize to a bullet bill

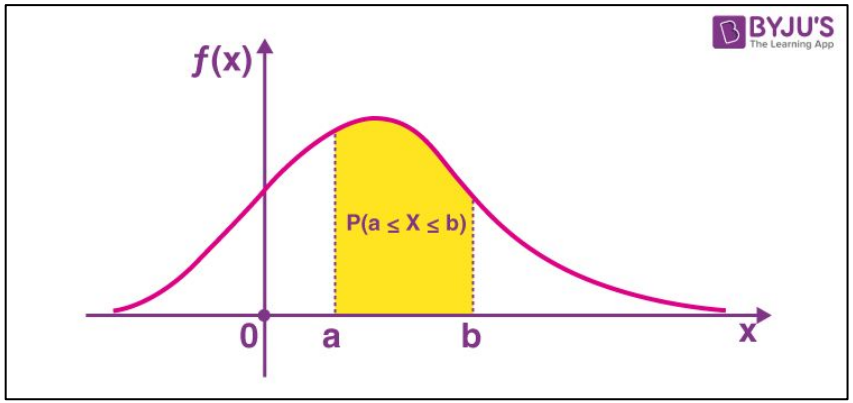


Probability Distributions

- **Probability distributions**: Functions that give probabilities of all possible **outcomes** for a **r.v.**
  - Intuitively, it describes a **r.v.** through its probabilities
  - We can learn a lot about a **r.v.** by its **probability distribution**
- For **discrete r.v.s**, we use **Probability Mass Functions (PMFs)**
  - $f(x) = P(X = x_i)$
  - "Probability of big X (r.v.) crystallizing to little x (a certain value)"
- For **continuous r.v.s**, we use **Probability Density Functions (PDFs)**
  - $f(x)$, where $P(a \leq X \leq b) = \int_a^b f(x)dx$
  - For continuous r.v.s, the probability of X crystallizing to a certain value is 0, so we're concerned with X crystallizing to any value within some interval

PDFs for Continuous Random Variables

- **Continuous r.v.s** are trickier because the probability X crystallizes to any one value is o
- **PDF**: f(x), where $P(a \leq X \leq b) = \int_a^b f(x)dx$
- Intuitively and visually, think of **PDF** as **a shape whose area represents probability**
  - Thus, the area of the entire shape is 1
  - f(x) evaluated at any certain point is NOT probability; here, probability is AREA



Special Types of Random Variables

- The really important types of **r.v.s** (which show up often) have names
- If your r.v. matches the "story" of a named random r.v., it makes your life easier
- X ~ Name(Value(s) of Key Parameter(s))
  - *Ex: $X \sim Bin(100, 0.10)$ is read as "X is distributed binomial with parameters 100 and 10"*
- **Parameters**: Named **r.v.s** are families, so **parameters** specify the **distribution** with a certain shape/center/spread
  - *Ex: $X \sim Bin(100, 0.10)$ is different from $Y \sim Bin(100, 0.50)$*

More on the "Mystery Box" Example...

- This **r.v.** can crystallize to any real number between 0 and 1 with equal probability
  - So this r.v. is "distributed **Unif(0, 1)**"
- That **r.v.** can crystallize to only 0 or 1, where it crystallizes to 1 with probability p; otherwise, it will crystallize to 0
  - So this r.v. is "distributed **Bern(p)**"

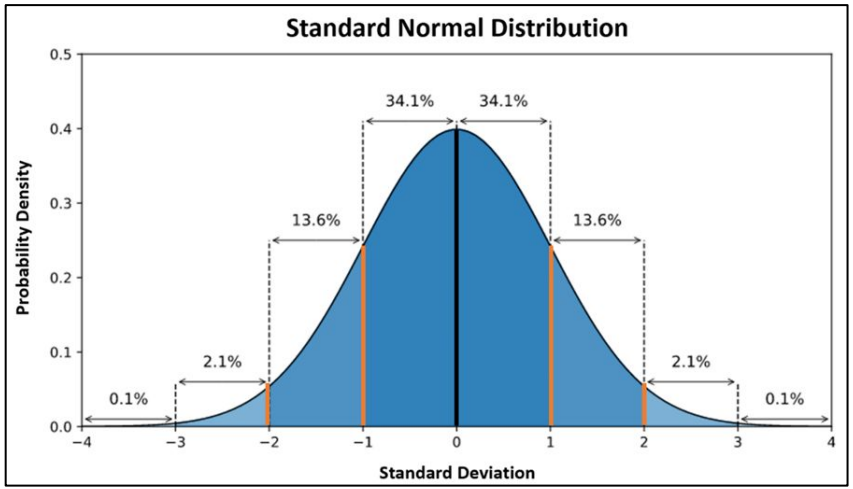
One More Thing...

- = and ~ are DIFFERENT
- **X = 1** says the r.v. X crystallizes to the value of 1
  - Recall this is a specific **event**, so we can calculate P(X = 1)
- **X ~ Bern(0.5)** says the r.v. X is distributed Bernoulli with p = 0.5

- Even if two r.v.s are **identically distributed**, they can still be different
  - *Ex: X ~ Bin(10, 0.5) and Y ~ Bin(10, 0.5) are identically distributed, but they can crystallize to different values*
  - *Imagine X counts the number of heads in 10 coin flips while Y counts the number of tails*

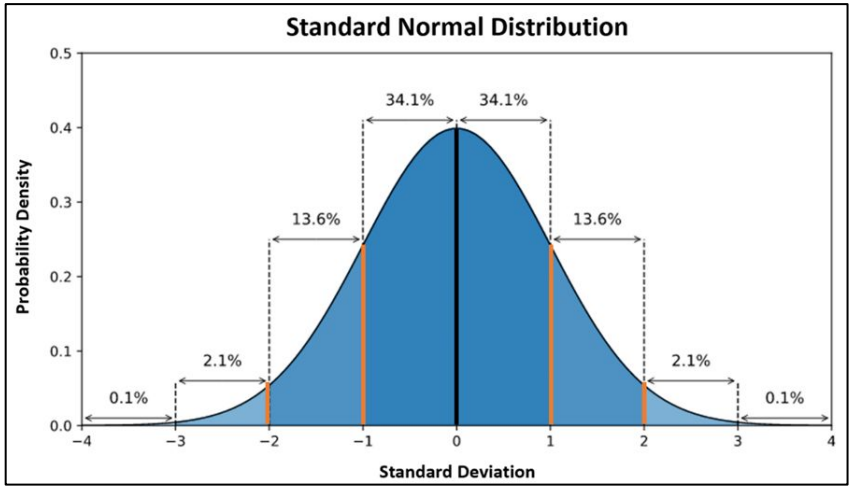
Normal Distribution

- **Normal distribution**: A symmetric and unimodal "bell shape" that approximates many **distributions**
- $N(\mu, \sigma)$ has 2 **parameters**
  - μ is **mean**
  - σ is **standard deviation**
- $Z(0, 1)$ is **Standard Normal**
  - 0 is **mean**
  - 1 is **standard deviation**



Standardizing and Z-Scores

- **Standardizing**: Transforming **normal r.v. (X)** into **standard normal r.v. (Z)**
  - Comparing in terms of **Z-scores** (**standard deviations**) is easier
- **Z-score**: Measure of how many **SDs** the **sample statistic** is away from its **mean**
  - Z-score = $(X - \mu) / \sigma$
  - Z-score for test statistic = (statistic - μ) / σ



The Important Functions for Normal Distribution

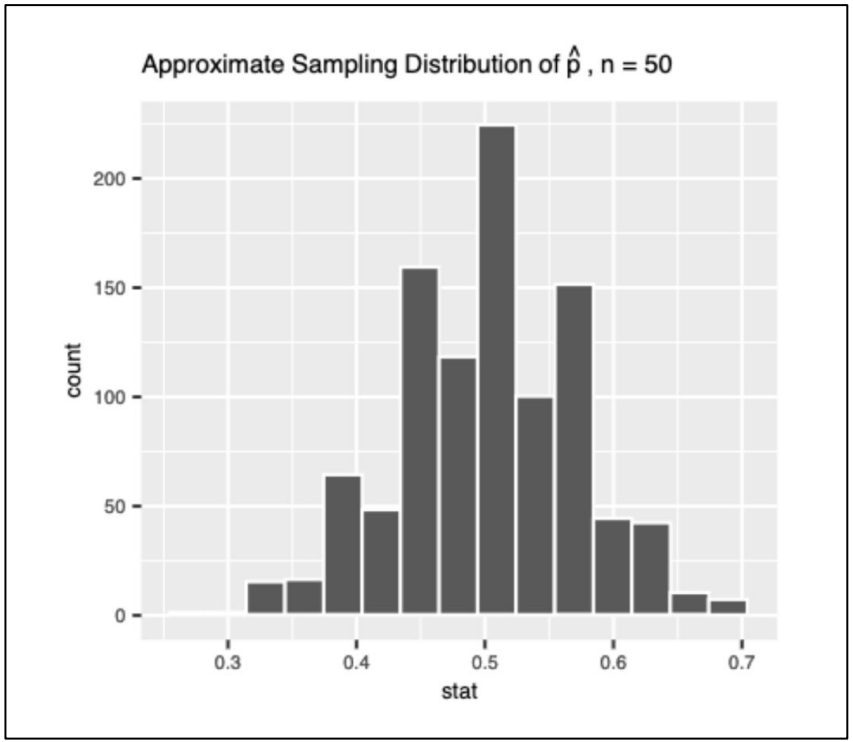
- **pnorm()**: Used to calculate **probabilities** on a **Normal distribution** (often, for **p-value** during **hypothesis test**)
  - *Ex: What is the **probability** a student scores an 1800 on the SAT if the scores are N(1500, 300)?*
- `pnorm(q = TEST-STAT, mean = MEAN, sd = STAN-DEV)`
  - *Ex: pnorm(q = 1800, mean = 1500, sd = 300) = 0.8413447*
- **qnorm()**: Used to calculate **quantiles** on a **Normal distribution** (often, for **critical value** during **confidence interval**)
  - *Ex: What score on the SAT would put a student in the 99th **quantile** (percentile)?*
- `qnorm(p = QUANTILE, mean = MEAN, sd = STAN-DEV)`
  - *Ex: qnorm(p = 0.99, mean = 1500, sd = 300) = 2197.904*

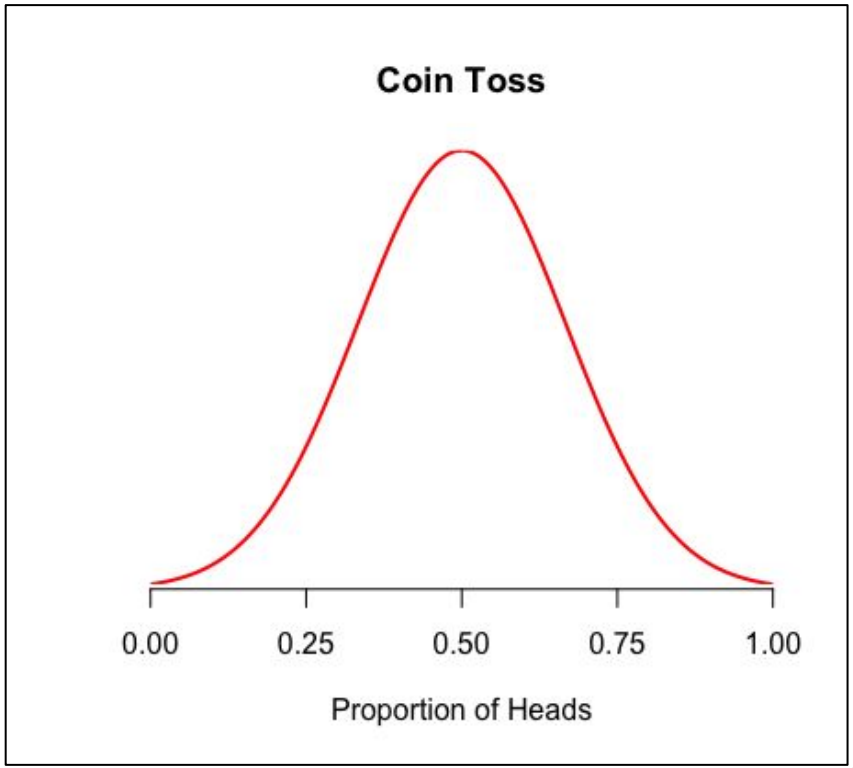
Why Does Any of This Matter?

- **Central Limit Theorem (CLT)**: For **random samples** and a **large sample size**, the **sampling distribution** of many **sample statistics** is approximately distributed **Normal**
  - Thus, when assumptions are met, we can conduct inference using the Normal distribution as a good approximation
  - We will revisit inference next week through this lens!

Content Review: Week 10

Null Distributions: Simulation-Based vs. Theory-based





A Visual Intuition for Central Limit Theorem

https://drive.google.com/file/d/128kvCSzPjRL7NMTtDRYlPAAY05RW7d3x/view?usp=drive_link

Theory-Based Inference

- Let's recast our **sample statistics** as **random variables**
- According to the **CLT**, when **assumptions** are met…
  - $\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$, where p = population proportion
  - $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, where μ = population mean and σ = population SD
- We often **standardize** our **sample statistic** to use **z-score** as our **test statistic**
  - This is because **Standard Normal dist.** is easy to use as our **Null dist.**
  - $\text{z-score} = \frac{X - \mu}{\sigma}$, where μ = population mean and σ = population SD

As a quick sanity check, why does it make sense to recast our sample statistics as random variables? Hint: Consider sampling variability and the “mystery box” intuition.

Question:

As a quick sanity check, why does it make sense to recast our sample statistics as random variables? Hint: Consider sampling variability and the "mystery box" intuition.

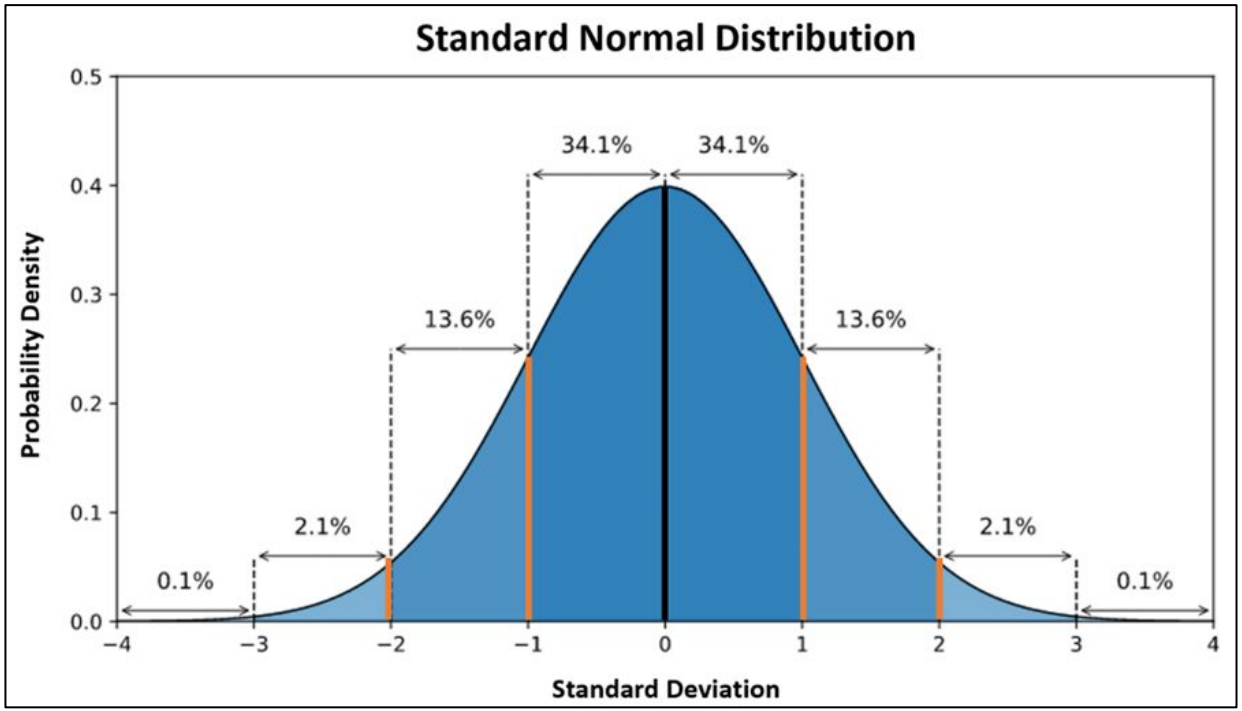
Due to **sampling variability**, **sample statistics** often differ from one another. For example, if I survey 400 people, my $\hat{p}$ would look different from yours if you surveyed 400 different people.

Thus, we can think of the **sample statistic** as a "mystery box" that will crystallize to a certain value after our sampling.

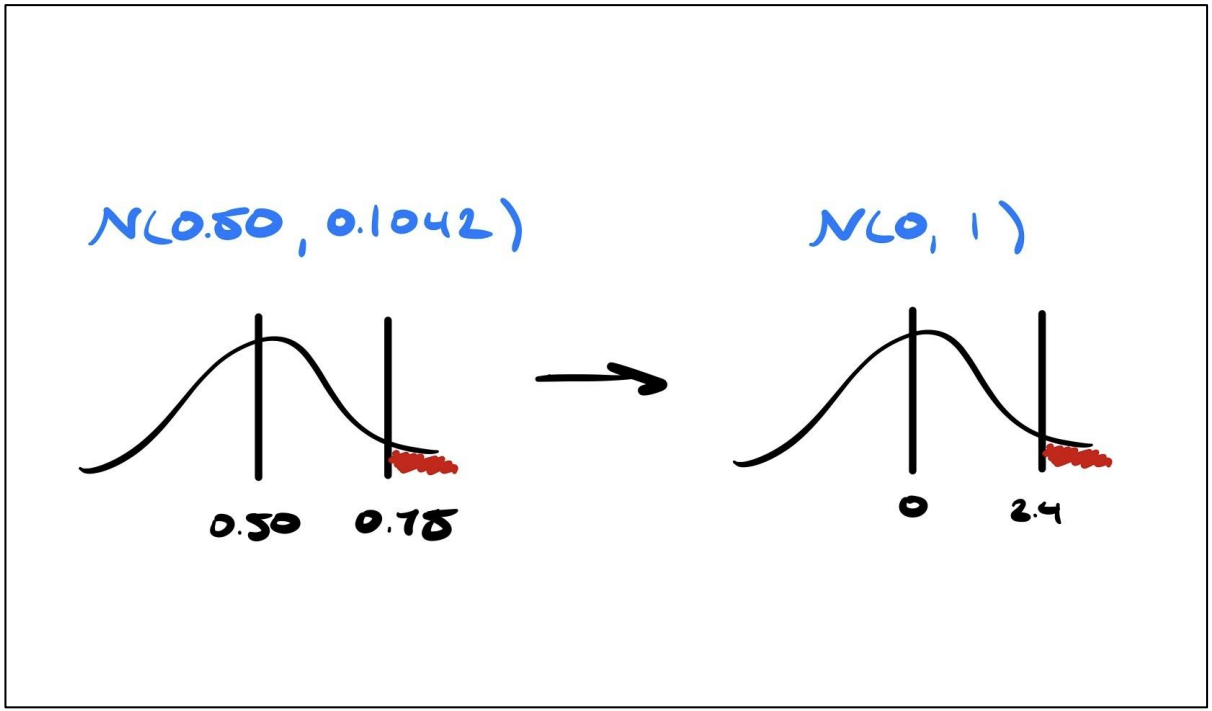
More on Test Statistic and Z-Score

- Up to now, we've been using our **(observed) sample statistic** as our **test statistic**
  - *"The prob. we get our observed test stat. of 75% heads (or more extreme) is..."*
- We can also use **z-score**, which is a standardized version of the **sample statistic**
  - *"The prob. we get a z-score of 2.4 (or more extreme) is..."*
  - It measures how many SDs the **sample statistic** is away from its **mean**
  - If **sample statistic ~ N(μ, σ)**, then **z-score ~ N(0, 1)** (**Standard Normal**)

Standard Normal Distribution

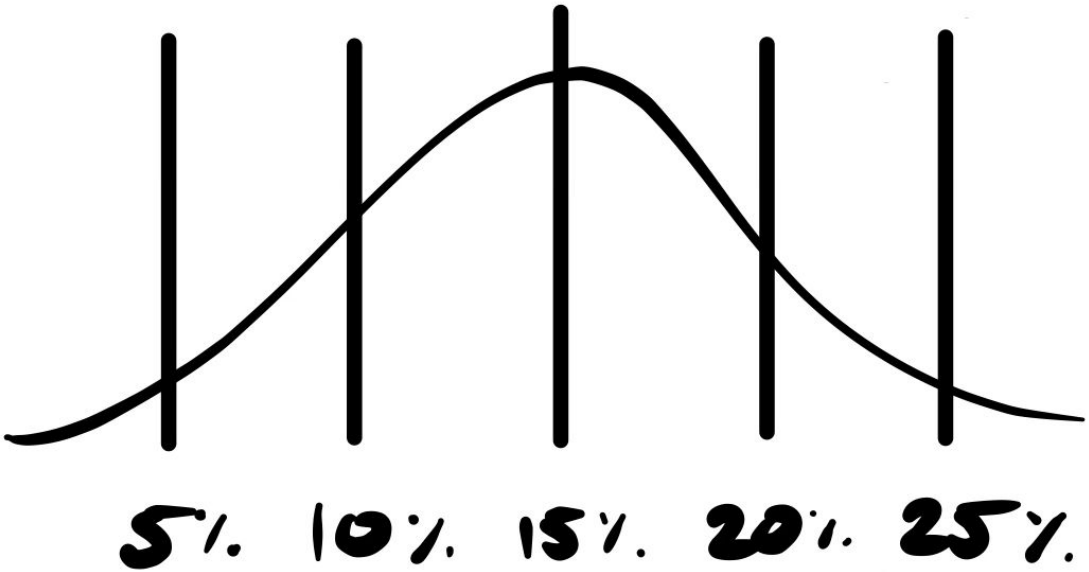


A Visual Intuition for Standardizing



If $\hat{p} \sim N(15\%, 5\%)$ and I get a sample with $\hat{p} = 25\%$, what is its z-score, and what does it mean?

$$\hat{p} \sim N(15\%, 5\%)$$



Question:

If $\hat{p} \sim N(15\%, 5\%)$ and I get a sample with $\hat{p} = 25\%$, what is its z-score, and what does it mean?

We're recasting our **sample statistic ($\hat{p}$)** as a **continuous r.v.**

We're given $\hat{p} \sim N(15\%, 5\%)$.
According to **CLT**, when assumptions are met, $X \sim N(\mu, \sigma)$.
Thus, mean = 15%, and SD = 5%.

z-score = $(X - \mu)/\sigma$, so z-score = 2.
We see 25% is 2 SDs away from 15%.

Theory-Based Hypothesis Tests (for Proportions)

- According to **CLT**, under the $\mathbf{H_o}$, $\hat{p} \sim N(p_0, \sqrt{\frac{p_0(1-p_0)}{n}})$
  - Remember $\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$
- Our **z-score (test statistic)** follows a **standard normal distribution**
  - $z \sim N(0, 1)$
- $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
  - Remember z-score = $(X - \mu)/\sigma$

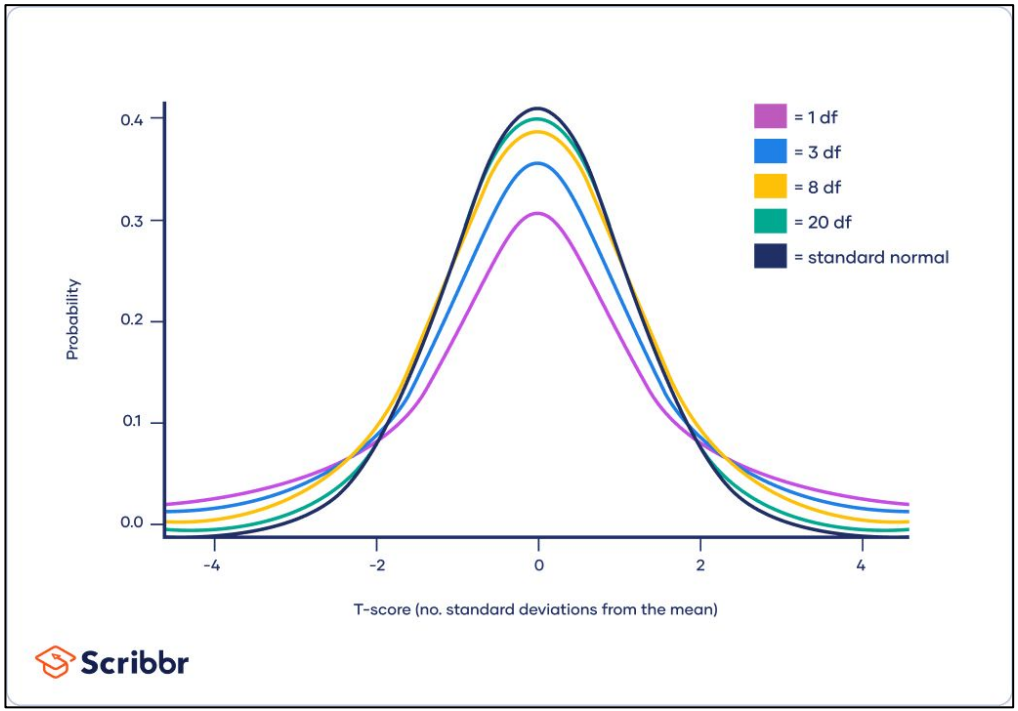
Theory-Based Confidence Intervals (for Proportions)

- A **CI** has the form of point estimate ± (critical value × SE)
  - **Critical value** is based on our desired **confidence level**
- According to **CLT**, $\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$
  - **SE** is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Thus, our **CI** (substituting in $\hat{\mathbf{p}}$ for **p**) is $\hat{p} \pm (z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$
  - **z*** is **critical value** in **norm. dist.**

For Means, We Have a Problem

- By **CLT**, $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, but we don't know **σ** (population SD), so we replace it with **s** (sample SD)
- When we use $\frac{s}{\sqrt{n}}$ as our SD, our **standardized test statistic** will follow a ***t* distribution** with df = n − 1 rather than **N(0, 1)**
  - Using the ***t* distribution** accounts for the extra variability introduced by using **s** as an estimate of **σ**
  - Our CI should be wider because we are now more uncertain

t distribution



For a t distribution, what happens as the degrees of freedoms increase?

Question:

For a t distribution, what happens as the degrees of freedoms increase?

As **degrees of freedom** increase for a ***t*** **distribution**, it looks more like a **normal distribution**.

Intuitively, as **degrees of freedom** increase, there is less uncertainty, so it becomes more appropriate to use **normal distribution**.

What Even Are Degrees of Freedom?

- **Degrees of freedom**: The number of values in the final calculation of a statistic that are free to vary
- With n = 3, if I tell you that $\bar{x} = 10$, $x_1 = 5$, $x_2 = 15$, then what must x_3 be? $x_3 = 10$!
- Thus, the is no variability/independence in that last observation, so degrees of freedom is n - 1

Theory-Based Hypothesis Tests (for Means)

- According to **CLT**, under the $\mathbf{H_o}$, $\bar{x} \sim N(\mu_0, \frac{\sigma}{\sqrt{n}})$
  - Remember we don't have **σ**, so we replace it with **s**
- Thus, $\bar{x} \sim N(\mu_0, \frac{s}{\sqrt{n}})$
- Now, our ***t*-score (standardized test statistic)** follows a ***t* distribution**
  - $t \sim t(\text{df} = n - 1)$
- $t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$
  - Remember z-score = $(X - \mu)/\sigma$... This is the ***t* distribution** analogue

Theory-Based Confidence Intervals (for Means)

- A **CI** has the form of point estimate ± (critical value × SE)
  - **Critical value** is based on our desired **confidence level**
- According to **CLT** and substituting in **s** for **σ**, $\bar{x} \sim N(\mu, \frac{s}{\sqrt{n}})$
  - **SE** is $\frac{s}{\sqrt{n}}$
- Thus, our **CI** is $\bar{x} \pm (t^* \times \frac{s}{\sqrt{n}})$
  - **t*** is critical value in ***t* distribution**

The Important Functions for Normal Distribution

- **pnorm()**: Used to calculate **probabilities** on a **normal distribution** (often, for **p-value** during **hypothesis test**)
  - *Ex: What is the **probability** a student scores an 1800 or less on the SAT if the scores are N(1500, 300)?*

```
- pnorm(q = TEST-STAT, mean = MEAN, sd = STAN-DEV)
```

  - *Ex: pnorm(q = 1800, mean = 1500, sd = 300) = 0.8413447*

- **qnorm()**: Used to calculate **quantiles** on a **normal distribution** (often, for **critical value** during **confidence interval**)
  - *Ex: What score on the SAT would put a student in the 99th **quantile** (percentile)?*

```
- qnorm(p = QUANTILE, mean = MEAN, sd = STAN-DEV)
```

  - *Ex: qnorm(p = 0.99, mean = 1500, sd = 300) = 2197.904*

The Important Functions for *t* distribution

- **pt()**: Used to calculate **probabilities** on a ***t* distribution** (often, for **p-value** during **hypothesis test**)
  - *Ex: What is the **probability** a student scores a 3 or less on an exam if the scores are ~ t(301 - 1)?*
- `pt(q = TEST-STAT, df = DEGREES-OF-FREEDOM)`
  - *`Ex: pt(q = 3, df = 301 - 1) = 0.9985369`*
- **qt()**: Used to calculate **quantiles** on a ***t* distribution** (often, for **critical value** during **confidence interval**)
  - *Ex: What score would put a student in the 99th **quantile** (percentile)?*
- `qt(p = QUANTILE, df = DEGREES-OF-FREEDOM)`
  - *`Ex: qt(p = 0.99, df = 301 - 1) = 2.338842`*

Let's Recap

- Want **probability**?
  - Use `pnorm()`, `pt()`
  - This is often done for **p-value** in **hypothesis testing**
- Want **quantile** (i.e. percentile)?
  - Use `qnorm()`, `qt()`
  - This is often done to find **z*** or **t*** in **confidence intervals**

Important Code for Theory-Based Inference

https://drive.google.com/file/d/1I2_ySaupN7crU8EwRVY1y_PFsfQP9nem/view?usp=drive_link

“Estimate” vs. “Statistic” in R

- **Estimate** is the **observed sample statistic** (i.e., the numeric quantity calculated with the data set)
  - *Here, the dataset had a sample correlation coefficient of -0.398*
- **Statistic** is the **standardized test statistic** (i.e., z-score or *t*-score)
  - *Here, that sample statistic is 7.07 standard errors below what we’d expect if the null hypothesis were true (i.e., if there is no correlation between age and vitamin D levels)*
  - *Here, the standardized test statistic is a t-score that’s distributed t(266)*

```
## # A tibble: 1 x 8
##   estimate statistic  p.value parameter conf.low conf.high method    alternative
##      <dbl>     <dbl>    <dbl>     <int>    <dbl>     <dbl> <chr>     <chr>
## 1   -0.398     -7.07 6.89e-12       266       -1    -0.309 Pearson'~ less
```

In the previous example, what values can “estimate” take on? What values can “statistic” take on?

Question:

In the previous example, what values can “estimate” take on? What values can “statistic” take on?

“Estimate,” as a sample correlation coefficient, can take on values in the interval $[-1, 1]$.

“Statistic,” as a t-score, can take on values in the interval $(-\infty, \infty)$.

Sample Size Calculation

- This is performed **before collecting data** to determine an appropriate **sample size** to gain desired **precision** for a **CI**
  - If my CI for average amount of sleep is between 1 and 23 hours, how helpful is that?
- **CI = point estimate ± (critical value × SE)**, where **margin of error = (critical value × SE)**
  - For proportions, **margin of error = z*** $\times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - For means, **margin of error = t*** $\times \frac{s}{\sqrt{n}}$
- We want our **margin of error** to be no larger than **B**, a bound
  - For proportions, $z* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq B \implies \frac{(z*)^2 \hat{p}(1-\hat{p})}{B^2} \leq n$
  - For means, $t* \times \frac{s}{\sqrt{n}} \leq B \implies \frac{(st*)^2}{B^2} \leq n$

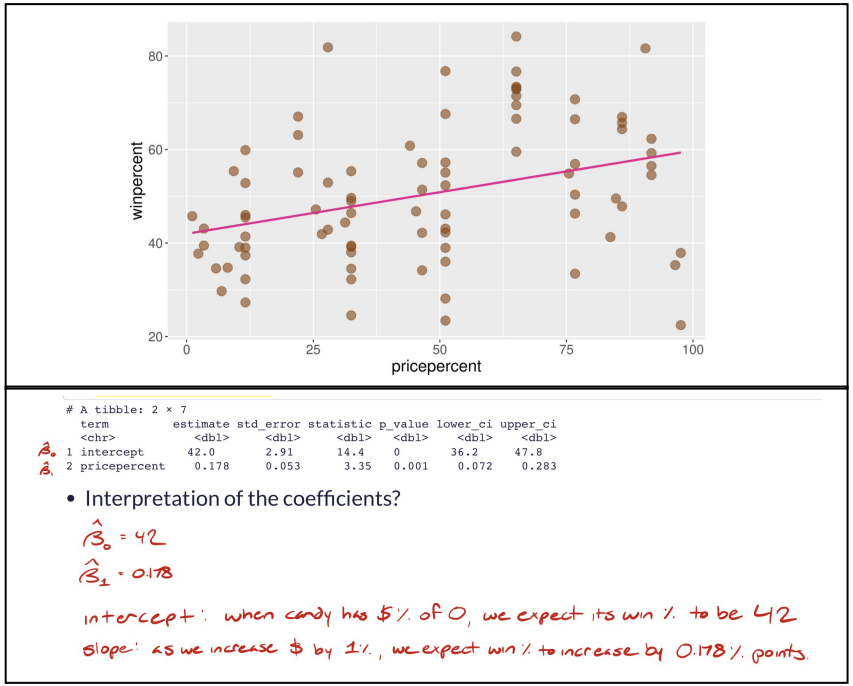
Content Review: Week 11

Linear Regression (In a Nutshell)

- **Linear regression**: Models the **linear** relationship between **numerical response variable (y)** and **explanatory variables (x)**, which can be either **numerical** or **categorical**
  - For now, we'll focus on **simple linear regression**, which only has one **explanatory variable**
- The form of this model is $\hat{y} = \hat{B}_0 + \hat{B}_1 x$
  - Note: $\hat{B}$ is supposed to represent beta hat ($\beta + \hat{}$)
- The **coefficients** ($\hat{B}_0$ **and** $\hat{B}_1$) have different interpretations depending on whether x is **numerical** or **categorical**

Explanatory Variable: Numerical

- When x is **numerical…**
  - The model represents a “line of best fit”
  - $\hat{B}_0$ is the **y-intercept**
    - *When price percentage equals 0%, the average win percentage is 42%*
  - $\hat{B}_1$ is the **slope**
    - *As price percentage increases by 1%, the win percentage increases by 0.178%, on average*
  - **Least-squares regression** finds the optimal values of $\hat{B}_0$ **and** $\hat{B}_1$ by minimizing **residuals** (errors)




```
# A tibble: 2 × 7
  term          estimate std_error statistic p_value lower_ci upper_ci
  <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept       42.0       2.91      14.4    0        36.2     47.8
2 pricepercent     0.178     0.053      3.35   0.001     0.072    0.283
```

$\hat{B}_0$
$\hat{B}_1$

- Interpretation of the coefficients?

$\hat{B}_0 = 42$

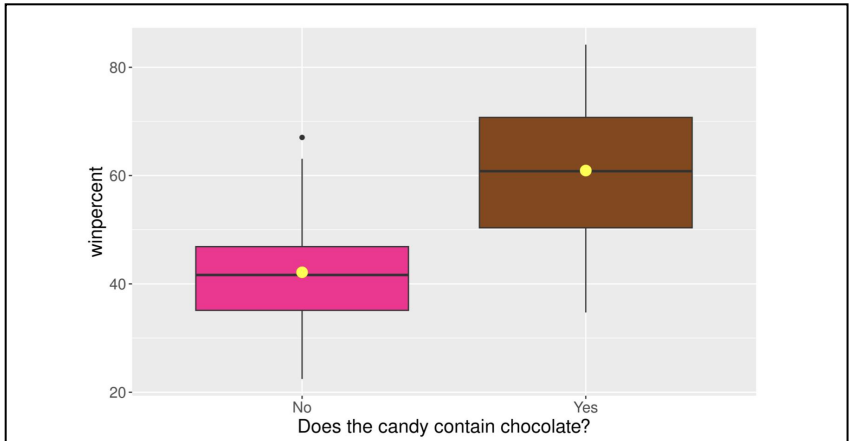
$\hat{B}_1 = 0.178$

intercept: when candy has $% of 0, we expect its win % to be 42

slope: as we increase $ by 1%, we expect win % to increase by 0.178% points

Explanatory Variable: Binary Categorical

- When x is **binary categorical**...
  - The model represents means (one for each of the two group)
  - $\hat{B}_0$ is the mean of y in the **baseline group** (when x = 0)
    - *For candy without chocolate, the average win percentage is 42.1%*
  - $\hat{B}_1$ is the **difference in means** of **other group** from **baseline group** ($\bar{y}_{other} - \bar{y}_{baseline}$)
    - *Candy with chocolate has a higher average win percentage than candy without chocolate by 18.8%*




```
# A tibble: 2 × 7
  term        estimate std_error statistic p_value lower_ci upper_ci
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept      42.1      1.65      25.6        0     38.9     45.4
2 chocolate      18.8      2.50       7.52       0     13.8     23.7
```

$\hat{B}_0$ → win % of non choc

$\hat{B}_1$ ↓

pos. slope bc. pos. relationship between choco. and win %.

Linear Regression: Code

- **Fitting the model**: Use this to build your model
  - `MODEL <- lm(Y-VAR ~ X-VAR, data = DATASET)`
  - *`model <- lm(winpercent ~ pricepercent, data = candy)`*
- **Getting the numbers**: Use this to summarize your model
  - `get_regression_table(MODEL)`
  - *`get_regression_table(model)`*
- **Predicting**: Use this for your model to predict y-value of new instances
  - `predict(MODEL, newdata = data.frame(Y-VAR = VALUE))`
  - *`predict(model, newdata = data.frame(pricepercent = 85))`*

More on Linear Regression

- **Interpolation**: Predicting values that fall **within** a dataset (generally good)
- **Extrapolation**: Predicting values that fall **outside** an observed range (generally not good)
- **Residual**: Error in **observed y** versus **predicted y** (**positive residual** means model **underestimated**; **negative residual** means model **overestimated**)
  - $\mathbf{e_i} = \mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}$ (observed - predicted)
- **Sample correlation coefficient (r)**: Measures **strength** of **linear relationship** between 2 **numeric variables** in a **sample**, ranging from -1 to 1
  - -1 is perfectly negative relationship
  - 1 is perfectly positive relationship

The model predicts a y-value of 26 while the (actual) observed y-value is 30. What is the residual, and what does it mean? Hint: e_i = observed - predicted.

Question:

The model predicts a y-value of 26 while the (actual) observed y-value is 30. What is the residual, and what does it mean? Hint: ei = observed - predicted.

$\mathbf{e_i} = \mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}$ (observed - predicted)

The **residual** is 4 (30 - 26). Thus, the model **underestimated** by 4.

Visually, the “line of best fit” is below the actual data point.

Population Model vs. Estimated Model

- **Population model**: $y = B_0 + B_1x + \varepsilon$
  - ε is **error/"random noise"** around the line (**population parameter** for the **residuals**)
  - $\varepsilon \sim N(0, \sigma)$
  - B_0 **and** B_1 are **population parameters**

- **Estimated model**: $\hat{y} = \hat{B}_0 + \hat{B}_1x$
  - This is what our "line of best fit" is
  - $\hat{B}_0$ **and** $\hat{B}_1$ are estimates of the **population parameters**
  - ε "disappears" because the **estimated model** is a straight line

Where else have we seen “hats” (^) used to indicate estimates?

Question:

Where else have we seen “hats” (^) used to indicate estimates?

Inference!

Recall $\hat{\mathbf{p}}$ (sample proportion) is used to estimate **p** (population proportion).

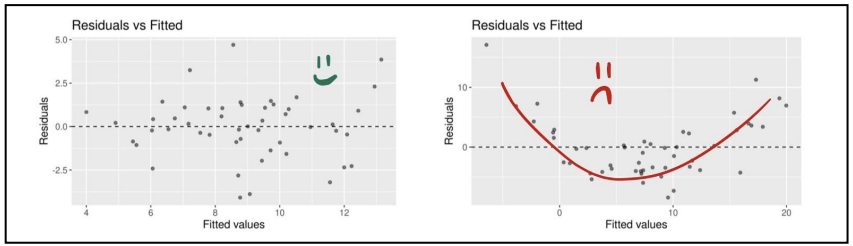
This is a common theme in statistics.

Assumptions for Linear Regression

- **Linearity**: The data shows a **linear** trend (thus, a linear model is appropriate)
- **Constant Variability**: The **variability** of the **response variable** about the line remains roughly constant as the **explanatory variable changes**
- **Independence**: Each observation is **independent** (i.e., value of one observation provide no information about value of others)
- **Normality**: The **residuals** (errors) are approximately **normally distributed**

Assumption #1: Linearity

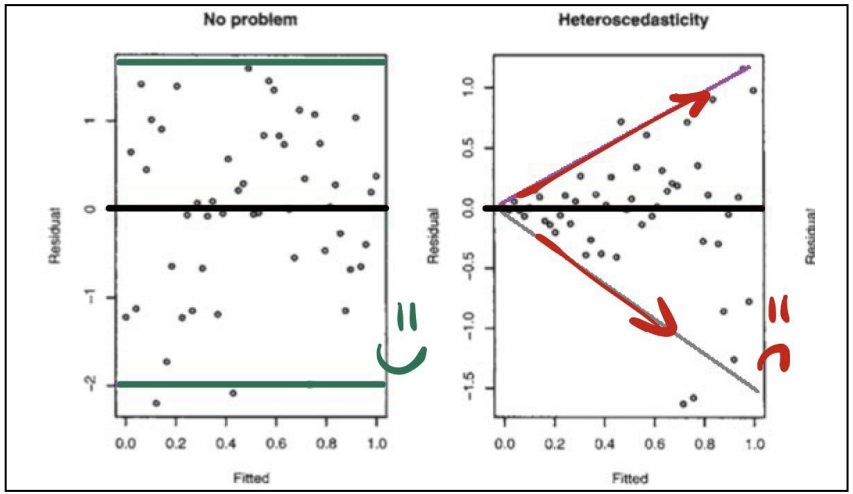
- Check via **residual plot**, which plots residuals of model across domain
- If data is linear, points should scatter from y = o randomly, with no pattern



- `ggplot(MODEL) + stat_fitted_resid()`
- *`ggplot(model) + stat_fitted_resid(alpha = 0.25)`*

Assumption #2: Constant Variance

- Check via **residual plot**, which plots residuals of model across domain
- Vertical spread of points should be roughly constant across domain, with no “fanning”
  - This interpretation is different from linearity; here, cite the upper and lower bounds (in green) to show there is no “fanning”



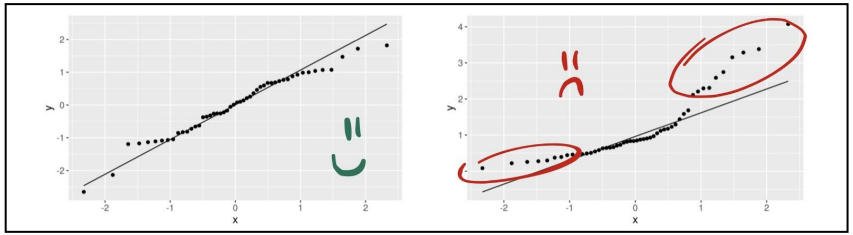
- `ggplot(MODEL) + stat_fitted_resid()`
- *`ggplot(model) + stat_fitted_resid(alpha = 0.25)`*

Assumption #3: Independence

- Check by considering **how data was collected**
- If there's **independence**, knowing observation #1 gives no information about observation #2
  - *Ex: If data was randomly sampled, then independence can be reasonably assumed*
  - *Ex: If data was collected within a family (and we're measuring blood sugar, e.g.), then independence might not apply. Why?*

Assumption #4: Normality

- Check via **Q-Q plot**, which plots residuals against theoretical quantiles of **normal distribution**
  - If residuals were perfectly **normally distributed**, they'd exactly follow the diagonal
  - We're not looking for perfect—just make sure it's reasonable
- Points should have a linear relationship, with no breaks at tails



- `ggplot(MODEL) + stat_normal_qq()`
- *`ggplot(model) + stat_normal_qq(alpha = 0.25)`*

Inference in Regression: Hypothesis Tests

- The **observed data** (x_i, y_i) is assumed to have been **randomly sampled** from a population where the **explanatory variable** (X) and the **response variable** (Y) follow a **population model**
  - **Population model**: $\mathbf{Y = B_0 + B_1X + \varepsilon}$
    - Like before, but we're now using capital letters to indicate **random variables**
  - **Estimated model**: $\mathbf{\hat{y} = \hat{B}_0 + \hat{B}_1x}$
- Usually, we're concerned with **slope parameter** (B_1)
  - $\mathbf{H_0: B_1 = 0}$ (i.e., the slope is zero, so there is no association between X and Y)
  - $\mathbf{H_A: B_1 \neq 0}$ (i.e., the slope is non-zero, so there is some association between X and Y)

Inference in Regression: Hypothesis Tests

- When **assumptions** are met (including 4 assumptions for linear regression), then the ***t* statistic** follows a ***t* distribution** with **degrees of freedom n − 2**, where n is the number of ordered pairs in the dataset
  - $\mathbf{t = (\hat{B}_1 - B_1^o)/SE(\hat{B}_1)}$
    - Recall our null hypothesis is (often) $\mathbf{B_1 = 0}$, so the $\mathbf{B_1^o}$ term can go away
  - $\mathbf{t = (\hat{B}_1)/SE(\hat{B}_1)}$
- Our computers can calculate this for us!
  - `get_regression_table(MODEL)`
  - *`get_regression_table(model)`*

Inference in Regression: Confidence Intervals

- **Confidence interval**: Recall the form of a confidence interval is **CI = sample statistic ± ME**
- **$CI = \hat{B}_1 \pm (t^* \times SE(\hat{B}_1))$**
  - **t^*** is the point on a ***t*** **distribution** with **n - 2 degrees of freedom** and **α/2 area to the right**
  - **"We are {α}% confident B_1 is in the CI; that is, with {α}% confidence, an increase in {explanatory variable} by 1 unit is associated with a change in average {response variable} between {lower bound} and {upper bound} units."**
  - *Ex: With 95% confidence, an increase in age of one year is associated with a change in average RFFT score between (-1.44, -1.08) points; i.e., a decrease in average RFFT score between 1.08 to 1.44 points.*
- Again, our computers can calculate this (use `get_regression_table()`)!

Confidence Interval vs. Prediction Interval

- **Confidence interval for mean response**: Tries to find plausible range for **parameter**
  - Centered at $\hat{y}$, with **smaller SE**
  - *Ex: We are 95% confident that the average RFFT score for individuals who are 50 years old is between 72.27 and 76.69 points.*

- **Prediction interval for individual response**: Tries to find plausible range for a **single, new observation**
  - Centered at $\hat{y}$, with **larger SE**
  - *Ex: For a 50-year-old individual, we predict, with 95% confidence, their RFFT score is between 28.87 and 120.10 points.*

Confidence Interval vs. Prediction Interval: Code

- OBSERVATION-OF-INTEREST <- data.frame(EXPL-VAR(S) = VALUE(S))
- predict(MODEL, newdata = OBSERVATION-OF-INTEREST, interval = "confidence", level = CONF-LEVEL)
  - *house_of_interest <- data.frame(livingArea = 1500, age = 20, bathrooms = 2, centralAir = "yes")*
  - *predict(model, house_of_interest, interval = "confidence", level = 0.95)*

- OBSERVATION-OF-INTEREST <- data.frame(EXPL-VAR(S) = VALUE(S))
- predict(MODEL, newdata = OBSERVATION-OF-INTEREST, interval = "prediction", level = CONF-LEVEL)
  - *house_of_interest <- data.frame(livingArea = 1500, age = 20, bathrooms = 2, centralAir = "yes")*
  - *predict(model, house_of_interest, interval = "prediction", level = 0.95)*

“Estimate” vs. “Statistic” in R

- **Estimate** is the **observed sample statistic** (i.e., the numeric quantity calculated with the data set)
  - *Here, $\hat{B}_1$ = **113**, so as living area increases by 1 unit, price increases by $113, on average*
- **Statistic** is the **standardized test statistic** (i.e., z-score or *t*-score)
  - *Here, **t** = **42.2**, so the sample statistic of $\hat{B}_1$ = 113 is 42.2 standard errors above what we’d expect if the null hypothesis were true (i.e., if β_1 = 0 so that there is no relationship between living area and price)*

```
## # A tibble: 2 x 7
##   term       estimate std_error statistic p_value lower_ci upper_ci
##   <chr>         <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    13439.    4992.       2.69   0.007    3648.   23231.
## 2 livingArea     113.      2.68     42.2    0         108.     118.
```

Content Review: Week 12

Introducing Multiple Linear Regression

- **Multiple linear regression**: Models the **linear** relationship between **numerical response variable (y)** and multiple **explanatory variables (x_1, x_2, ..., x_p)**, which can be either **numerical** or **categorical**
- The form of this model is $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1 + ... + \hat{B}_p x_p$
  - Note: $\hat{B}$ is supposed to represent beta hat ($\beta + \hat{}$)
- $\hat{B}_k$ (coefficient of **predictor x_k**) is predicted mean change in **y (response variable)** corresponding to **1 unit change in x_k** when **all other predictors are held constant**
  - If x_k is **numerical**, think of slope
  - If x_k is **categorical**, think of difference in means (of group where $x_k = 1$ from baseline group)

For houses, if I want to predict price based on living area and whether or not there's central air, what is p (number of predictors)?

Question:

For houses, if I want to predict price based on living area and whether or not there's central air, what is p (number of predictors)?

We'll use linear regression to model this relationship.

$\hat{y}$ = price

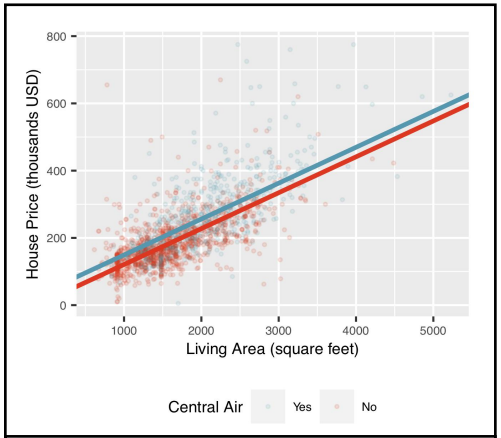
x_1 = living area (numerical)

x_2 = whether or not there's central air (categorical)

Thus, p = 2.

Example: Houses

- **Variables**: price ($\hat{y}$), living area (x_1), whether or not there's central air (x_2)
  - x_1 is **numerical**, x_2 is **categorical**
  - **Baseline group** is houses WITH central air
- **Estimated model**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1 + \hat{B}_2 x_2$
  - **Line when $x_2 = 0$ (houses WITH central air)**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1$
    - **y-intercept** = $\hat{B}_0$, **slope** = $\hat{B}_1$
  - **Line when $x_2 = 1$ (houses WITHOUT central air)**: $\hat{y} = (\hat{B}_0 + \hat{B}_2) + \hat{B}_1 x_1$
    - **y-intercept** = $\hat{B}_0 + \hat{B}_2$, **slope** = $\hat{B}_1$




```
## # A tibble: 3 x 2
##   term            estimate
##   <chr>              <dbl>
## 1 intercept         42595.   β̂0
## 2 livingArea          107.   β̂1
## 3 centralAir: No   -28451.   β̂2
```

when living area, x_1, is held constant, this is the average difference in houseprice, y

Example: Houses

- **Variables**: price ($\hat{y}$), living area (x_1), whether or not there's central air (x_2)
  - x_1 is **numerical**, x_2 is **categorical**
  - **Baseline group** is houses WITH central air
- **Estimated model**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1 + \hat{B}_2 x_2$
  - **Line when $x_2 = 0$ (houses WITH central air)**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1$
    - **y-intercept** = $\hat{B}_0$, **slope** = $\hat{B}_1$
  - **Line when $x_2 = 1$ (houses WITHOUT central air)**: $\hat{y} = (\hat{B}_0 + \hat{B}_2) + \hat{B}_1 x_1$
    - **y-intercept** = $\hat{B}_0 + \hat{B}_2$, **slope** = $\hat{B}_1$

- Since we have **multiple variables**, be careful interpreting the **coefficients**
  - **$\hat{B}_0$**: For **houses with central air ($x_2 = 0$)**, when **living area (x_1)** equals 0, the **price ($\hat{y}$)** is **\$42,595 ($\hat{B}_0$)**, on average
  - **$\hat{B}_1$**: Controlling for **central air (x_2)**, as **living area (x_1)** increases by 1 unit, **price ($\hat{y}$)** increases by **\$107 ($\hat{B}_1$)**, on average
  - **$\hat{B}_2$**: Controlling for **living area (x_1)**, **houses without central air ($x_2 = 0$)** cost **\$28,451 ($\hat{B}_2$)** less than **houses with central air ($x_2 = 1$)**, on average

The General "Formulas" for Equal-Slopes (When x_2 Is Categorical)

- $\hat{B}_0$ is **y-intercept of line when $x_2 = 0$**
  - *Ex: For houses with central air ($x_2 = 0$), when living area (x_1) equals 0, the price ($\hat{y}$) is \$42,595 ($\hat{B}_0$), on average*
- Since this is equal-slopes, $\hat{B}_1$ is **slope of both lines** (a.k.a. increase in $\hat{y}$ after 1-unit increase in x_1, **controlling for x_2**)
  - *Ex: Controlling for central air (x_2), as living area (x_1) increases by 1 unit, price ($\hat{y}$) increases by \$107 ($\hat{B}_1$), on average*
- $\hat{B}_0 + \hat{B}_2$ is y-intercept of line $x_2 = 1$, so $\hat{B}_2$ is **difference in $\hat{y}$ between both lines ($\hat{y}_{other} - \hat{y}_{baseline}$), controlling for x_1**
  - *Ex: Controlling for living area (x_1), houses without central air ($x_2 = 0$) cost \$28,451 ($\hat{B}_2$) less than houses with central air ($x_2 = 1$), on average*

Looking at the tibble, how can we tell what's the baseline group?

Question:

Looking at the tibble, how can we tell what's the baseline group?

Remember the **baseline group** is when $x_k = 0$ for some categorical predictor x_k.

Things are relative to the **baseline group**, so the tibble presents the "change" with the **categorical predictor** (to $x_k = 1$ from $x_k = 0$).

Thus, the **baseline group** is the OPPOSITE of the group shown.

Baseline Group

```
## # A tibble: 3 x 2
##   term             estimate
##   <chr>               <dbl>
## 1 intercept         42595.   β̂0
## 2 livingArea          107.   β̂1
## 3 centralAir: No   -28451.   β̂2
```

when living area, x_1, is held constant, this is the average difference in houseprice, y

The output tells us "centralAir: No" has an estimate of -28,451. Thus, "centralAir: Yes" (a.k.a. houses WITH central air) is our baseline group.

Categorical Variables with 2+ Categories

- **Linear regression** can accommodate **categorical variables** with 2+ categories
  - *Ex: We can predict RFFT score with the categorical variable of education, which can be "Lower Secondary," "Higher Secondary," or "University"*
- When **x** is a **categorical variable** with k + 1 categories…
  - $\hat{\mathbf{B}}_0$ represents the **mean of y** in the **baseline group** (one of those k + 1 categories)
  - $\hat{\mathbf{B}}_k$ represents the **difference in means**—specifically, going from x = 0 (**baseline group**) to x = k (one of the other groups)
  - Thus, $\hat{\mathbf{B}}_k = \bar{\mathbf{y}}_{\text{group k}} - \bar{\mathbf{y}}_{\text{baseline}}$
- We can confirm our answers with some data wrangling
- Let's look at an example…

Interpreting a categorical predictor with several levels

$$\widehat{RFFT} = 40.9 + 14.8(Edu_{LS}) + 32.1(Edu_{HS}) + 45.0(Edu_{Univ})$$

- When x is a categorical variable with $k+1$ levels...
  - $\hat{\beta}_0$ represents the mean of y in the baseline group
  - $\hat{\beta}_k$ represents the difference in means; specifically, going from $x = 0$ to $x = k$
- Mean RFFT score is 40.9 points among those with at most a Primary education.
- The mean RFFT score among those with at most a University education is 45 points higher than those with at most a Primary education: $40.9 + 45 = 85.9$ points.
- The `Edu_new: Univ` coefficient equals $\overline{y}_{Univ} - \overline{y}_{Primary} = 45$

```
prevend.samp %>%
  group_by(Edu_new) %>%
  summarize(mean_RFFT = mean(RFFT))
```

```
## # A tibble: 4 x 2
##   Edu_new     mean_RFFT
##   <fct>           <dbl>
## 1 Primary          40.9
## 2 Lower Sec        55.7
## 3 Higher Sec       73.1
## 4 Univ             85.9
```

+14.8, +32.2, +45

```
model <- lm(RFFT ~ Edu_new,
            data = prevend.samp)
get_regression_table(model) %>%
  select(term, estimate)
```

```
## # A tibble: 4 x 2
##   term                  estimate
##   <chr>                    <dbl>
## 1 intercept                40.9
## 2 Edu_new: Lower Sec       14.8
## 3 Edu_new: Higher Sec      32.1
## 4 Edu_new: Univ            45.0
```

estimate : diff in means

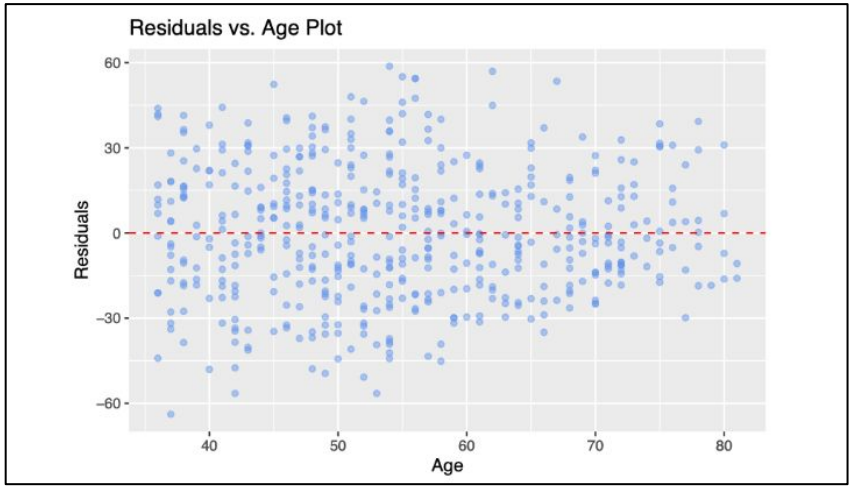
$\hat{\beta}_0 = 40.9$, $\hat{\beta}_1 = 14.8$, $\hat{\beta}_2 = 32.1$, $\hat{\beta}_3 = 45.0$

Assumptions for (Multiple) Linear Regression

- **Linearity**: For each **predictor variable x_k** , the change in the **predictor** is **linearly related** to change in the **response variable** when the values of **all other predictors** are held constant
- **Constant Variability**: The **residuals** (errors) have approximately **constant variance**
- **Independence**: Each observation is **independent** (i.e., value of one observation provide no information about value of others)
- **Normality**: The **residuals** (errors) are approximately **normally distributed**

Assumption #1: Linearity

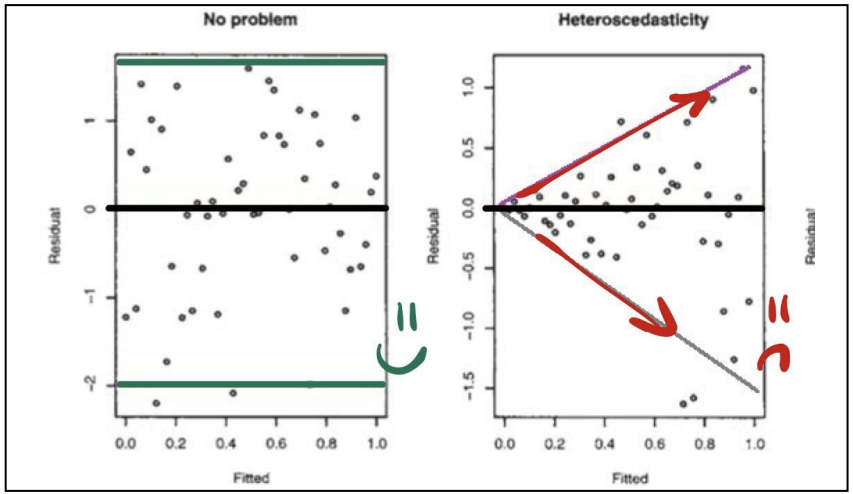
- Check via "**residual vs. predictor" plot with ggplot()**
  - For each **numerical predictor**, plot the **residuals** on the y-axis and the **predictor values** on the x-axis
- If data is linear, points should scatter from y = 0 randomly, with no pattern



- `ggplot(MODEL, aes(y = .resid, x = NUM-PREDICTOR) + geom_point() + geom_hline(yintercept = 0)`
- *Ex: `ggplot(mod_rfft, aes(y = .resid, x = Age)) + geom_point(alpha = 0.5, col = "cornflowerblue") + geom_hline(yintercept = 0, lty = 2, col = "red") + labs(y = "Residuals", x = "Age", title = "Residuals vs. Age Plot")`*

Assumption #2: Constant Variability

- Check via **residual plot**, which plots residuals of model across domain
- Vertical spread of points should be roughly constant across domain, with no “fanning”
  - This interpretation is different from **linearity**; here, cite the upper and lower bounds (in green) to show there is no “fanning”



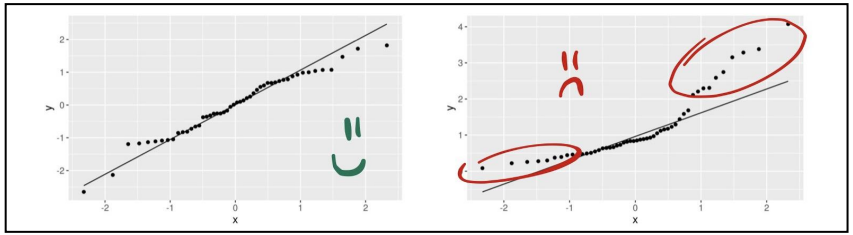
- `ggplot(MODEL) + stat_fitted_resid()`
- *`Ex: ggplot(model) + stat_fitted_resid(alpha = 0.25)`*

Assumption #3: Independence

- Check by considering **how data was collected**
- If there's **independence**, knowing observation #1 gives no information about observation #2
  - *Ex: If data was randomly sampled, then independence can be reasonably assumed*
  - *Ex: If data was collected within a family (and we're measuring blood sugar, e.g.), then independence might not apply. Why?*

Assumption #4: Normality

- Check via **Q-Q plot**, which plots residuals against theoretical quantiles of **normal distribution**
  - If residuals were perfectly **normally distributed**, they'd exactly follow the diagonal
  - We're not looking for perfect—just make sure it's reasonable
- Points should have a linear relationship, with no breaks at tails



- `ggplot(MODEL) + stat_normal_qq()`
- *Ex: `ggplot(model) + stat_normal_qq(alpha = 0.25)`*

Returning to Inference: Population Model vs. Estimated Model

- **Population model**: $y = B_0 + B_1x_1 + \ldots + B_px_p + \varepsilon$
  - **ε** is **error/"random noise"** around the line (**population parameter** for the **residuals**)
  - **$\varepsilon \sim N(0, \sigma)$**
  - **B_k** is **population parameter**

- **Estimated model**: $\hat{y} = \hat{B}_0 + \hat{B}_1x_1 + \ldots + \hat{B}_px_p$
  - This is what our "line of best fit" is
  - **$\hat{B}_k$** is estimate of the **population parameter**
  - **ε** "disappears" because the **estimated model** is a straight line

Inference in (Multiple) Regression: Hypothesis Tests

- The **observed data** is assumed to have been **randomly sampled** from a population where the **explanatory variable** (X) and the **response variable** (Y) follow a **population model**
  - **Population model**: $\mathbf{Y = B_0 + B_1X_1 + \ldots + B_pX_p + \varepsilon}$
    - Like before, but we're now using capital letters to indicate **random variables**
  - **Estimated model**: $\mathbf{\hat{y} = \hat{B}_0 + \hat{B}_1x_1 + \ldots + \hat{B}_px_p}$
- Usually, we're concerned with **slope parameter** (B_k)
  - $\mathbf{H_0: B_k = 0}$ (i.e., there is no association between X_k and Y after controlling for all other predictors in the model)
  - $\mathbf{H_A: B_k \neq 0}$ (i.e., there is an association between X_k and Y after controlling for all other predictors in the model)

Inference in (Multiple) Regression: Hypothesis Tests

- When **assumptions** are met (including 4 assumptions for multiple linear regression), then the **t-statistic** follows a **t-distribution** with **degrees of freedom n − p - 1**, where n is the number of cases and p is the number of predictors
  - $\mathbf{t = (\hat{B}_k - B_k^0)/SE(\hat{B}_k)}$
    - Recall our null hypothesis is (often) $\mathbf{B_k = 0}$, so the $\mathbf{B_k^0}$ term can go away
  - $\mathbf{t = (\hat{B}_k)/SE(\hat{B}_k)}$
- Our computers can calculate this for us!
  - `get_regression_table(MODEL)`
  - *`Ex: get_regression_table(model)`*

Inference in (Multiple) Regression: Confidence Intervals

- **Confidence interval**: Recall the form of a confidence interval is **CI = sample statistic ± ME**
- **$CI = \hat{B}_k \pm (t^* \times SE(\hat{B}_k))$**
  - **t^*** is the point on a **t-distribution** with **$n-p-1$ degrees of freedom** and **$\alpha/2$ area to the right**
  - **"With {α}% confidence, an increase in {explanatory variable} by 1 unit is associated with a change in average {response variable} between {lower bound} and {upper bound} units when holding {other explanatory variables in model} constant."**
  - *Ex: With 95% confidence, statin users have an average RFFT score that is between 4.2 points lower to 5.9 points higher than non statin users when holding age constant. Here, x_k is categorical, so this is better interpreted as a difference in means.*
- Again, our computers can calculate this for us (use `get_regression_table()`)!

Confidence Interval vs. Prediction Interval

- **Confidence interval for mean response**: Tries to find plausible range for **parameter**
  - Centered at $\hat{y}$, with **smaller SE**
  - *Ex: We are 95% confident that the average price of 20 year-old, 1,500 square-feet Saratoga houses with central air and 2 bathrooms is between $199,919 and $211,834*
- **Prediction interval for individual response**: Tries to find plausible range for a **single, new observation**
  - Centered at $\hat{y}$, with **larger SE**
  - *Ex: For a 20 year-old, 1,500 square-foot Saratoga house with central air and 2 bathrooms, we predict, with 95% confidence, the price will be between $73,885 and $337,869*

Confidence Interval vs. Prediction Interval: Code

- `OBSERVATION-OF-INTEREST <- data.frame(EXPL-VAR(S) = VALUE(S))`
- `predict(MODEL, newdata = OBSERVATION-OF-INTEREST, interval = "confidence", level = CONF-LEVEL)`
  - *Ex: house_of_interest <- data.frame(livingArea = 1500, age = 20, bathrooms = 2, centralAir = "yes")*
  - *predict(model, house_of_interest, interval = "confidence", level = 0.95)*

- `OBSERVATION-OF-INTEREST <- data.frame(EXPL-VAR(S) = VALUE(S))`
- `predict(MODEL, newdata = OBSERVATION-OF-INTEREST, interval = "prediction", level = CONF-LEVEL)`
  - *Ex: house_of_interest <- data.frame(livingArea = 1500, age = 20, bathrooms = 2, centralAir = "yes")*
  - *predict(model, house_of_interest, interval = "prediction", level = 0.95)*

Two Types of Mult. Linear Regression: Equal-Slopes, Varying-Slopes

- **Equal-Slopes**: Assumes **change in y** associated with change in **1 explanatory variable**—a.k.a. the slope—DOES NOT DEPEND on **other explanatory variable(s)** in model
  - Visually, we see equal slopes in the lines
- **Estimated model**: $\hat{y} = \hat{B}_0 + \hat{B}_1x_1 + \hat{B}_2x_2 + \ldots + \hat{B}_px_p$
  - We see there are no terms where the x variables interact with each other
- Code: `— <- lm(— ~ — + —, data = —)`

- **Varying-slopes model**: Assumes **change in y** associated with change in **1 explanatory variable**—a.k.a. the slope—DOES DEPEND on **other explanatory variable(s)** in model, so interaction term(s) is present
  - Visually, we see different slopes in the lines
- **Estimated model**: $\hat{y} = \hat{B}_0 + \hat{B}_1x_1 + \hat{B}_2x_2 + \hat{B}_3x_1x_2 + \ldots + \hat{B}_px_p$
  - We see there is an interaction term between x_1 and x_2: $\mathbf{\hat{B}_3x_1x_2}$
- Code: `— <- lm(— ~ — * —, data = —)`

For houses, if I want to predict price based on living area and whether or not there's central air—now with a varying slopes model—what is p (number of predictors)?

Question:

For houses, if I want to predict price based on living area and whether or not there's central air—now with a varying slopes model—what is p (number of predictors)?

We'll use linear regression (with varying-slopes) to model this relationship.

$\hat{y}$ = price

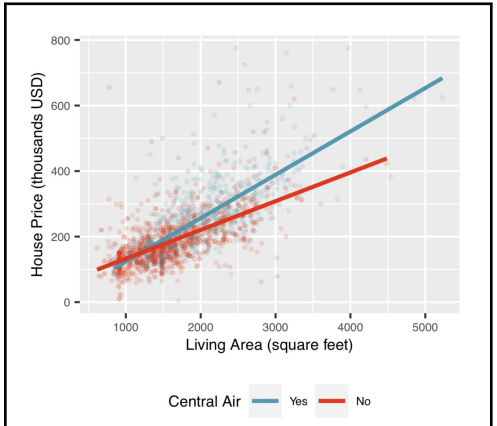
x_1 = living area (numerical)

x_2 = whether or not there's central air (categorical)

Thus, p = 2—like last time!

Example: Houses (But with Varying-Slopes)

- **Variables**: price ($\hat{y}$), living area (x_1), whether or not there's central air (x_2)
  - x_1 is **numerical**, x_2 is **categorical**
  - **Baseline group** is houses WITH central air
- **Estimated model**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1 + \hat{B}_2 x_2 + \hat{B}_3 x_1 x_2$
  - **Line when $x_2 = 0$ (houses WITH central air)**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1$
    - **y-intercept** = $\hat{B}_0$, **slope** = $\hat{B}_1$
  - **Line when $x_2 = 1$ (houses WITHOUT central air)**: $\hat{y} = (\hat{B}_0 + \hat{B}_2) + (\hat{B}_1 + \hat{B}_3)x_1$
    - **y-intercept** = $\hat{B}_0 + \hat{B}_2$, **slope** = $\hat{B}_1 + \hat{B}_3$
    - Notice the **slopes** are different!




```
## # A tibble: 4 x 2
##   term                          estimate
##   <chr>                            <dbl>
## 1 intercept                      -8248.   B̂0
## 2 livingArea                       132.   B̂1
## 3 centralAir: No                 53226.   B̂2
## 4 livingArea:centralAirNo          -44.6  B̂3
```

difference in intercepts

coefficient on interaction term

Example: Houses (But with Varying-Slopes)

- **Variables**: price ($\hat{y}$), living area (x_1), whether or not there's central air (x_2)
  - x_1 is **numerical**, x_2 is **categorical**
  - **Baseline group** is houses WITH central air
- **Estimated model**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1 + \hat{B}_2 x_2 + \hat{B}_3 x_1 x_2$
  - **Line when $x_2 = 0$ (houses WITH central air)**: $\hat{y} = \hat{B}_0 + \hat{B}_1 x_1$
    - **y-intercept** = $\hat{B}_0$, **slope** = $\hat{B}_1$
  - **Line when $x_2 = 1$ (houses WITHOUT central air)**: $\hat{y} = (\hat{B}_0 + \hat{B}_2) + (\hat{B}_1 + \hat{B}_3)x_1$
    - **y-intercept** = $\hat{B}_0 + \hat{B}_2$, **slope** = $\hat{B}_1 + \hat{B}_3$
    - Notice the **slopes** are different!

- $\hat{B}_0$: For **houses with central air ($x_2 = 0$)**, when **living area (x_1)** equals 0, the **price ($\hat{y}$)** is **-$8,248 ($\hat{B}_0$)**, on average
- $\hat{B}_1$: For **houses with central air ($x_2 = 0$)**, as **living area (x_1)** increases by 1 unit, **price ($\hat{y}$)** increases by **$132 ($\hat{B}_1$)**, on average
- $\hat{B}_2$: When **living area (x_1)** equals 0, **houses without central air ($x_2 = 1$)** cost **$53,226 ($\hat{B}_2$)** more than **houses with central air ($x_2 = 0$)**, on average
- $\hat{B}_3$: **Houses without central air ($x_2 = 1$)** have a lower slope than houses with central air by **$44.6/unit ($\hat{B}_3$)**. For houses without central air **($x_2 = 1$)**, as **living area (x_1)** increases by 1 unit, **price ($\hat{y}$)** increases by **$87.4 ($\hat{B}_1 - \hat{B}_3$)**, on average

The General "Formulas" for Varying-Slopes (When x_2 Is Categorical)

- $\hat{B}_0$ is **y-intercept of line when $x_2 = 0$**
  - *Ex: For houses with central air ($x_2 = 0$), when living area (x_1) equals 0, the price ($\hat{y}$) is -$8,248 ($\hat{B}_0$), on average*
- $\hat{B}_1$ is **slope of line when $x_2 = 0$**
  - *Ex: For houses with central air ($x_2 = 0$), as living area (x_1) increases by 1 unit, price ($\hat{y}$) increases by $132 ($\hat{B}_1$), on average*
- $\hat{B}_0 + \hat{B}_2$ is y-intercept of line when $x_2 = 1$ (houses without central air), so $\hat{B}_2$ is **difference in y-intercepts between both lines ($b_{other} - b_{baseline}$)**
  - *Ex: When living area (x_1) equals 0, houses without central air ($x_2 = 1$) cost $53,226 ($\hat{B}_2$) more than houses with central air ($x_2 = 0$), on average*
- $\hat{B}_1 + \hat{B}_3$ is slope of line when $x_2 = 1$ (houses without central air), so $\hat{B}_3$ is **difference in slopes between both lines ($m_{other} - m_{baseline}$)**
  - *Ex: Houses without central air ($x_2 = 1$) have a lower slope than houses with central air by $44.6/unit ($\hat{B}_3$)*

Inference with Varying-Slopes

- Same idea as before, but now we can infer about **population interaction coefficient** (B_3) instead of **population slope coefficient** (B_1)
  - H_0: $B_3 = 0$ (i.e., association/slope between y and x_1 doesn't differ by category)
  - H_A: $B_3 \neq 0$ (i.e., association/slope between y and x_1 differs by category)
- Again, our computers give us this info with `get_regression_table()`!

When should I use equal-slopes vs. varying-slopes?

Question:

When should I use equal-slopes vs. varying-slopes?

Consider your goal with the model.

With varying-slopes, certain questions (like the average difference in cholesterol between diabetic groups, controlling for age) can't be answered.

With equal-slopes, certain questions (like whether or not the relationship/slope differs between groups) can't be answered.

If r ranges from -1 to 1, what are the possible values for r^2?

Question:

If r ranges from -1 to 1, what are the possible values for r^2?

0-1!

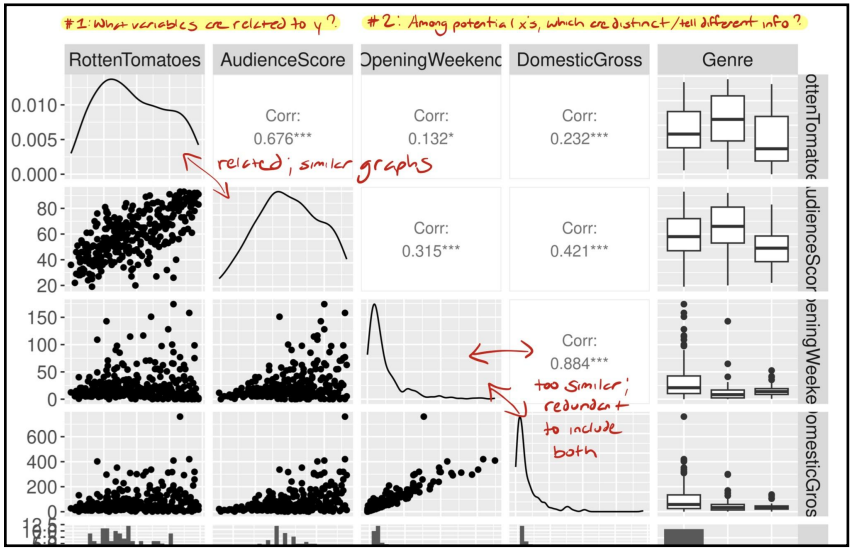
As a result of squaring the numbers, r^2 can only take on non-negative values.

r^2: Coefficient of Determination

- **r^2**: Percent of **total variation** in y (**response variable**) explained by the **model**
  - **$r^2 = (r)^2 = Var(\hat{y}_i)/Var(y_i)$**
  - If the **linear model** perfectly captured the **variability** in the observed data, then $Var(\hat{y}_i) = Var(y_i)$; thus, **$r^2$** would be 1
  - If **r^2** is too low, try different model; however, **r^2** only increases as new **predictors** are added to a model
- **adj(r^2)**: Value of **r^2** adjusted for size of model (penalizes too-large models)
  - **$adj(r^2) = r^2 \times ((n - 1)/(n - p - 1))$**
  - n is sample size, p is number of predictors in model
- Basically, graph your data and pick the model with **highest adj(r^2)**
  - `glance(MODEL)`
  - *`Ex: glance(model)`*

Model Building Guidance

- In addition to looking at **adj(r^2)**, consider your **explanatory variables** in the **model**
  - You want them to **explain different aspects** of the **response variable**
  - It would be redundant to have both RottenTomatoes and AudienceScore in a model, for example
- Use `ggpair()` to see relationship between multiple **explanatory variables**
  - If the graphs look alike, this tells you the **variables** are similar—consider removing one of them



Content Review: Week 13

Big Picture Overview

- We introduce 3 (+2) more tools in our inference toolkit
- These are extensions of things we've seen before
  - Paired t-test
  - ANOVA
  - Chi-squared
- Also…
  - Fisher's exact test
  - Effect size in 2x2 tables

An Introduction to Pairing

- Two-sample numerical data can be **paired** or **unpaired** (i.e., independent)
- Thus far, we've been working with **unpaired**
  - Observations cannot be matched on a one-to-one
  - *Ex: Considering SAT scores for students who studied versus students who did not, we can't match Alice, who studied, with Bob, who didn't—they're completely different people!*
- Now, let's consider studies with **paired** measurements
  - Each observation can be logically matched to another observation in the data
  - *Ex: Considering SAT scores for a group of 10 students before and after studying, we're matching Alice's old score with her new score*

If we want to measure the effect of new wetsuits on swimmers, should we have paired data or unpaired data?

Question:

If we want to measure the effect of new wetsuits on swimmers, should we have paired data or unpaired data?

While both strategies could work, this research question might be best answered with a **paired study**.

It'd be better to keep our swimmers consistent (since everyone has their own velocity, generally). Thus, our data can be paired "before and after."

Paired t-test: Example

- 12 swimmers had their velocity measured using an (old) swimsuit and using a (new) wetsuit
  - These are paired data (e.g., swimmer 1 swimsuit can be matched with swimmer 1 wetsuit)
- Conducting a **non-paired t-test** (what we've done before), we get $\bar{x}_{swimsuit} - \bar{x}_{wetsuit} = 0.0775$ m/s, with a p-value of 0.18
- For **paired t-test**, we look at **đ**, the **sample mean** of differences in velocities
  - If swimmer 1 swam 1.5 m/s with wetsuit and 1.4 m/s with swimsuit, their difference is 0.1 m/s
  - **đ** would be average of 12 differences (e..g,
- **δ** is the **population mean** of difference in velocities (theoretically, for all swimmers—not just 12)

Paired t-test: Example

- **H_o: $\delta = 0$**, the **population mean** difference in swim velocities between swimming with a swimsuit versus a wetsuit equals 0
  - That is, wetsuits do NOT change swim velocities
- **H_A: $\delta \neq 0$**, the **population mean** difference in swim velocities between swimming with a swimsuit versus a wetsuit is non-zero
  - That is, wetsuits DO change swim velocities
- $t = (\bar{d} - \delta_o)/(s_d/\sqrt{n})$, where **t** is our **standardized test statistic** (***t*-score**)
- $t \sim t(df = n - 1)$, where **n** is number of differences/pairs
- $95\%\ CI = \bar{d} \pm (t^* \times s_d/\sqrt{n})$, where **t*** is point on $t(df = n - 1)$ that has area 0.025 to its right (assuming $\alpha = 0.05$)

Paired t-test: Code

- As always, our computers do the math for us—we just need to code and interpret!
- Use `t_test()`
  - We're used to this from the tidyverse
  - A paired *t*-test is just a single-mean test on the differences

Paired t-Test: Code for t_test()

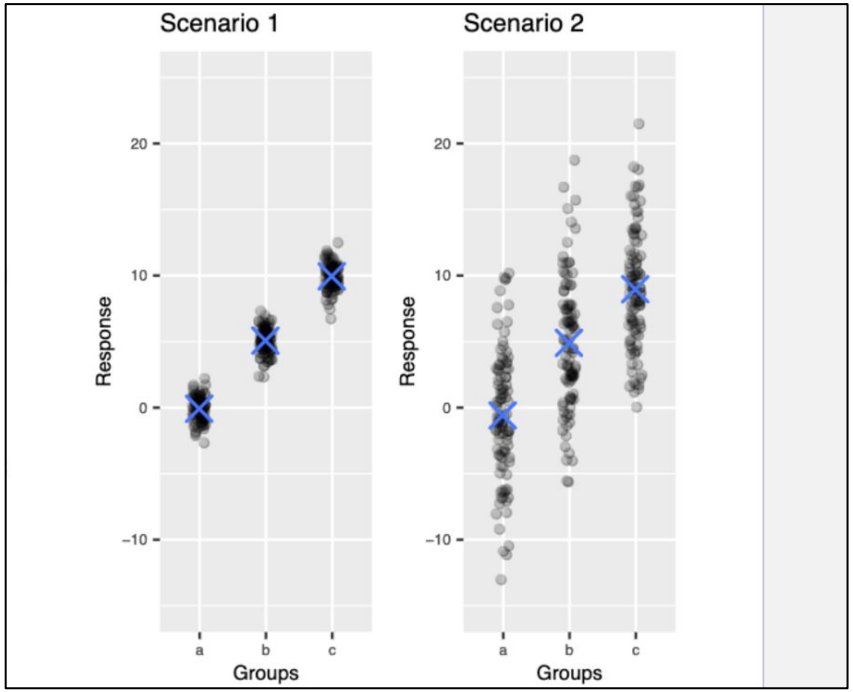
- **General form**: `DATASET %>% t_test(response = RESPONSE-VAR.diff)`
  - *`swim %>% t_test(response = velocity.diff)`*
  - Again, very similar to before, but now we're adding "**.diff**" because we're interested in the difference for each pair
- **Hypothesis tests**: `DATASET %>% t_test(response = RESPONSE-VAR.diff) %>% select(statistic, p_value, estimate)`
  - *`swim %>% t_test(response = velocity.diff) %>% select(statistic, p_value, estimate)`*
- **Confidence intervals**: `DATASET %>% t_test(response = RESPONSE-VAR.diff) %>% select(lower_ci, upper_ci)`
  - *`swim %>% t_test(response = velocity.diff) %>% select(lower_ci, upper_ci)`*

ANOVA: Analysis of Variance

- **ANOVA**: Test for when **response variable** is **numerical** and **explanatory variable** is **categorical (with more than 2 categories)**
  - H_o: $\mu_1 = \mu_2 = ... = \mu_k$ (i.e., variables are independent)
  - H_A: At least 1 mean is not equal to the rest (i.e., variables are dependent)
- Test statistic is **F-statistic**
  - **F = standardardized variance BETWEEN groups / standardized variance WITHIN groups**
- If H_o is true, **F-statistic** should be roughly equal to 1 (variance between groups should be equal to variance within groups)
- If H_A is true, **F-statistic** should be larger than 1
  - *Ex: If F is 3.88, the variance BETWEEN groups is 3.88 times larger than the variance WITHIN groups, which suggests the population means are different*

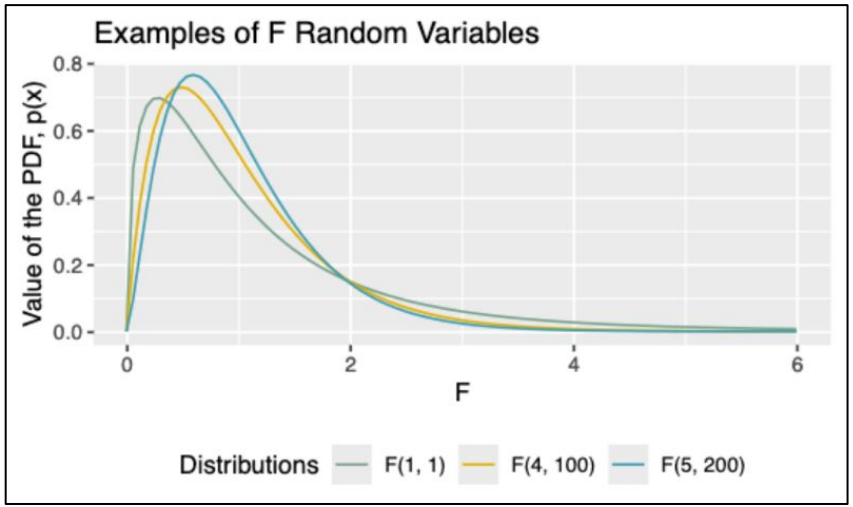
ANOVA: Intuition

- **Scenario 1**, there is little variability WITHIN groups but much more variability BETWEEN groups
  - It's plausible these groups come from different populations
- **Scenario 2**, there is a lot of variability WITHIN groups, so we're less sure… this would correspond to a low **F-statistic**



ANOVA: Theory-Based Inference

- When the ANOVA assumptions (next few slides) are satisfied, the **F-statistic** follows an **F distribution**, with two degrees of freedom: $\mathbf{df_1}$ and $\mathbf{df_2}$
- That is, **F-statistic ~ F($\mathbf{df_1}$, $\mathbf{df_2}$)**
  - $\mathbf{df_1 = n_{groups} - 1}$, $\mathbf{df_2 = n_{observations} - n_{groups}}$
- The p-value is P(F > observed F-statistic)—area to the RIGHT

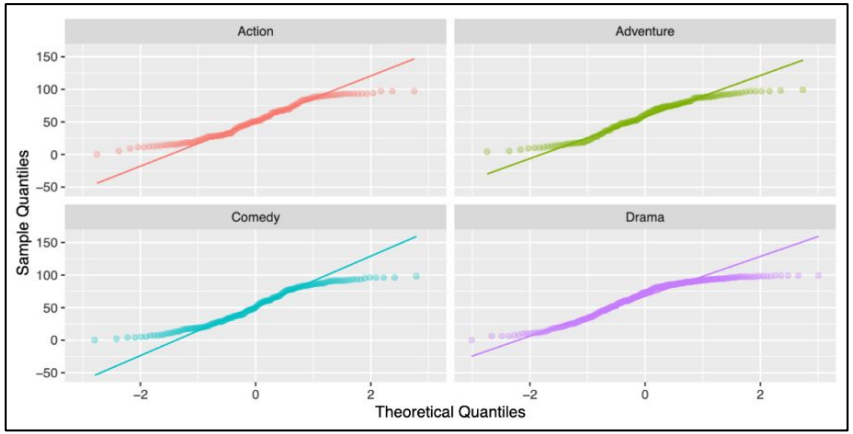


Assumptions for (Theory-Based) ANOVA

- **Assumption #1**: Observations are independent within and across groups
  - Think about study design/context (i.e., read the description)
- **Assumption #2**: Data within each group are approximately **normal**
  - Use **Normal Q-Q plots** (if data are perfectly normal, they follow the line in the Q-Q plot exactly)
  - As **sample size** increases, deviation from normality becomes less of a concern
- **Assumption #3**: **Variability** across groups is about equal
  - The rule of thumb is we want to see **largest variance / smallest variance < 3**, which we can find via data wrangling

Assumption #2: Normality

- Check via **Q-Q plot**, which plots residuals against theoretical quantiles of **normal distribution**
  - If residuals were perfectly **normally distributed**, they'd exactly follow the diagonal
  - We're not looking for perfect—just make sure it's reasonable
- Points should have a linear relationship, with no breaks at tails




```
ggplot(movies_subset, aes(sample = RottenTomatoes, col = Genre)) + geom_qq(alpha = 0.30) + stat_qq_line() + facet_wrap(~ Genre) + labs(y = "Sample Quantiles", x = "Theoretical Quantiles") + guides(col = "none")
```

Assumption #3: Constant Variability

- Check via **data wrangling**
- Remember **variance** is a measure of variability
- We don't expect the **variances** to be exactly the same across groups
  - As a rule of thumb, we want the ratio of largest variance to smallest variance to be less than 3
  - That is, **largest variance / smallest variance < 3**

```
## # A tibble: 4 x 3
##   Genre       var     n
##   <chr>     <dbl> <int>
## 1 Action     724.   170
## 2 Adventure  734.   163
## 3 Comedy     800.   191
## 4 Drama      680.   384
```

```
800/680
```

```
## [1] 1.176471
```

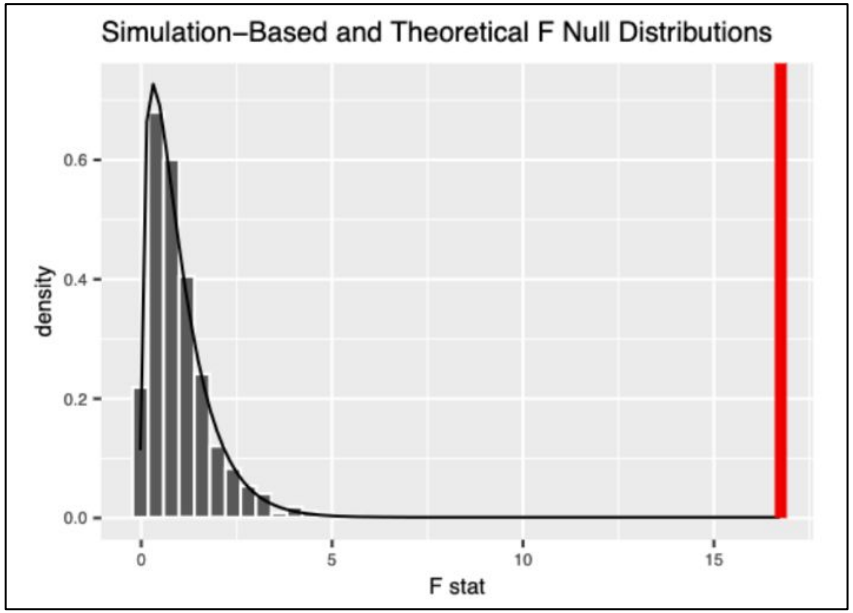
```
movies_subset %>% drop_na(Genre,
RottenTomatoes) %>% group_by(Genre) %>%
summarize(var = var(RottenTomatoes), n = n())
```

ANOVA: Code

- **Strategy #1**: Tidyverse R
  - `ANOVA_MODEL <- anova(lm(Y-VAR ~ X-VAR, data = DATASET))`
  - `tidy(ANOVA_MODEL)`
    - *`movies_anova <- anova(lm(RottenTomatoes ~ Genre, data = movies_subset))`*
    - *`tidy(movies_anova)`*
- **Strategy #2**: Base R
  - `ANOVA_MODEL <- aov(Y-VAR ~ X-VAR, data = DATASET)`
  - `summary(ANOVA_MODEL)`
    - *`movies_anova <- aov(RottenTomatoes ~ Genre, data = movies_subset)`*
    - *`summary(movies_anova)`*

ANOVA: More Intuition

- Remember, when assumptions are met, **F-statistic ~ F(df1, df2)**
- Remember, under $\mathbf{H_0}$, **F-statistic** should be equal to 1
- If **F-statistic** is much higher than 1, the variance between groups is much larger than the variance within groups, suggesting the $\mathbf{H_A}$
- "More extreme" is to the RIGHT!



ANOVA: Afterwards, How Do We Know Which Group Is Different?

- After seeing evidence against $\mathbf{H_o}$ (i.e., 1 of the means is different), how do we see which group is different?
- We'll conduct pairwise t-tests (like what we've been doing before)
- To keep Type I errors in check, we use adjusted alpha level, **α***
  - If we don't, the probability of a Type I error explodes as we do many pairwise t-tests!
- **$\alpha^* = \alpha/K$** , where **K is the total number of possible two-way comparisons**
  - **$K = k(k-1)/2$**, where **k is the total number of groups**
  - *Ex: If α = 0.05, then when there are 4 groups, α* = 0.05/6 = 0.0083*
- Our computers can calculate **α*** for us (the "bonferroni" correction)

ANOVA: Afterwards, Pairwise t-Tests

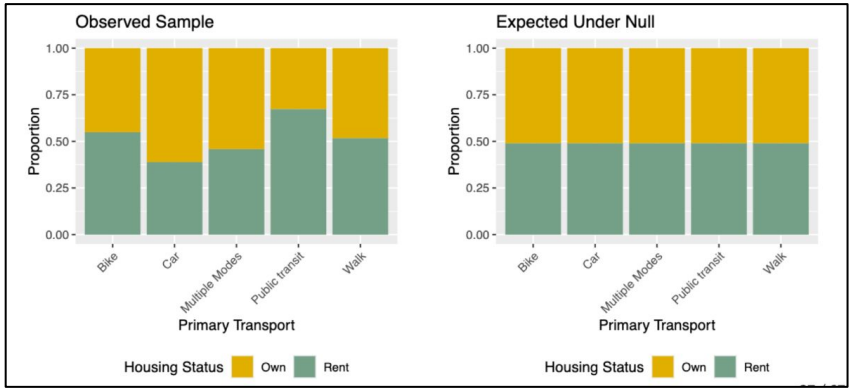
- **Pairwise t-Tests** isn't in tidyverse R, so we're using base R (with different syntax)!
- Remember to use "bonf" if you want to computer to calculate **α***
  - `pairwise.t.test(DATASET$Y-VAR, DATASET$X-VAR, p.adjust.method = "bonf")`
  - *`pairwise.t.test(movies_subset$RottenTomatoes, movies_subset$Genre, p.adjust.method = "bonf")`*

Chi-Squared

- **Chi-Squared test**: Test for when both **response variable** and **explanatory variable** are **categorical**, and **at least one has more than 2 categories**
  - $\mathbf{H_0}$: The variables are independent
  - $\mathbf{H_A}$: The variables are dependent
- If **response variable** and **explanatory variable** were both binary categorical, we'd just use **difference in proportions** (like before)!
- Our test statistic is χ^2 (which, essentially, sums and squares every z-score so that negatives are accounted for)
  - $\chi^2 = \Sigma(\text{observed} - \text{expected} / \sqrt{\text{expected}})^2$
- The intuition is best explained by graphs and tables...

Chi-Squared: Intuition

- If **variables were actually independent** (i.e., primary transport doesn't affect housing status), then we'd expect a graph to look like the one on the right
- We'd also **expect** certain values
  - We expect (0.0389)(1534) = 59.72 residents who rent homes and use a bike, but we observe 67 residents
  - **Chi-squared** squares and sums these values to see "total extremity"



If observed = expected, what would χ^2 equal? Hint: $\chi^2 = \Sigma(\text{observed} - \text{expected} / \sqrt{\text{expected}})^2$.

Question:

If observed = expected, what would χ^2 equal? Hint: $\chi^2 = \Sigma(\text{observed} - \text{expected} / \sqrt{\text{expected}})^2$.

If **observed** = **expected**, then $(\text{observed} - \text{expected} / \sqrt{\text{expected}})^2 = 0$, so we just sum of a bunch of zeros to get $\boldsymbol{\chi^2 = 0}$.

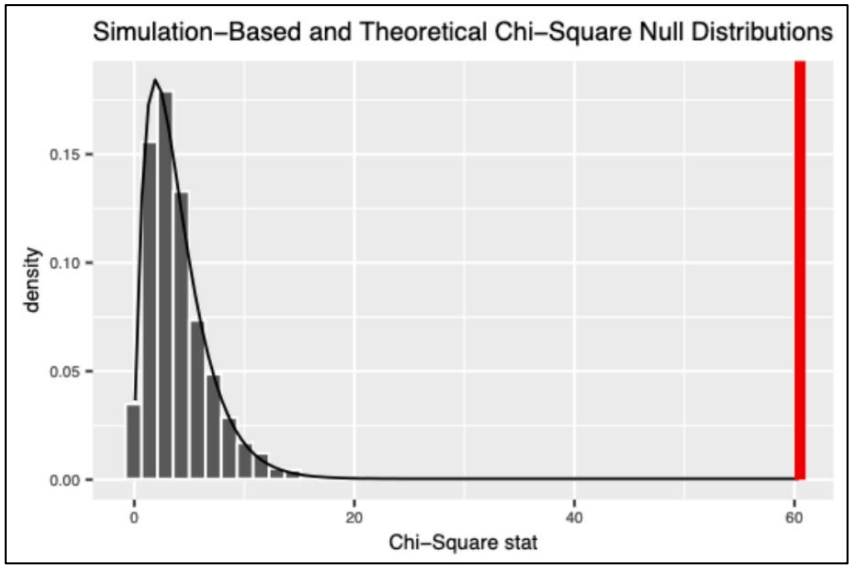
This is why we fail to reject the null hypothesis if $\boldsymbol{\chi^2}$ **is near 0**—that implies what we observed is very close to what we expected under the null.

Assumptions for Chi-Squared

- **Assumption #1**: Random sampling
- **Assumption #2**: There are at least 10 observations in each cell (check via **data wrangling**)
  - `count(DATASET, X-VAR, Y-VAR)`
  - *`count(grammar, Education, oxford_comma)`*
- These assumptions must be met for the test statistic to be approximately distributed χ^2 with degrees of freedom $(r-1)(c-1)$, where **r** is the number of rows and **c** is the number of columns

Chi-Square: Intuition

- Our test statistic is χ^2, which essentially sums and squares every (standardized) difference between what we expect and what we observe
  - $\chi^2 = \Sigma(\text{observed} - \text{expected} / \sqrt{\text{expected}})^2$
- $\chi^2 \sim \chi^2(df = (r - 1)(c - 1))$
  - r = number of rows
  - c = number of columns
- χ^2 quantifies how far the observed results deviate from what is expected under H_o
  - A larger value shows stronger evidence against H_o of independence (thus, "more extreme" is to the RIGHT!)



Chi-Squared: Code

- **Strategy #1**: Tidyverse R
  - `chisq_test(DATASET, Y-VAR ~ X-VAR)`
  - *`chisq_test(somerville, housing ~ primary_transport)`*
- **Strategy #2**: Base R
  - `chisq.test(DATASET$X-VAR, somerville$Y-VAR)`
  - *`chisq.test(somerville$primary_transport, somerville$housing)`*

Chi-Square: Afterwards, Examining Residuals

- We could compare the **observed** versus **expected values** to identify which table cells are contributing the most to the **test statistic**
- Instead of having to look back and forth between two tables, look at the table of **residuals**
- **Residuals** with a **large magnitude** contribute the most to the χ^2 **statistic**
  - If a **residual** is **positive**, the observed value is greater than the expected value
  - If a **residual** is **negative**, the observed value is less than the expected

Chi-Square: Afterwards, Examining Residuals: Code

- **General form**: `chisq.test(DATASET$X-VAR, DATASET$Y-VAR)$residuals`
  - *chisq.test(somerville$primary_transport, somerville$housing)$residuals*

Recap: Inference Scenarios and Test Statistics

https://drive.google.com/file/d/1i1XTclseg1_CPu6eECBuaoKDy6XuNMyn/view?usp=drive_link

Questions?

Problem Solving Strategies and Common Mistakes

First, Load All Relevant Libraries

- `library(tidyverse)`
- `library(infer)`
- `library(ggplot2)`
- `library(gglm)`
- `library(moderndive)`
- `library(dplyr)`
- `library(broom)`
- `library(knitr)`
- There may be more I'm forgetting… it doesn't hurt to load more to be safe!

Correctly Use the "P-Value Formula"

- **"If {null hypothesis} were true, then the probability of observing {test statistic} or {more extreme} would be {p-value}."**
  - This is "interpreting the p-value"
- **"Because {p-value} is a {high/low} probability compared to {alpha}, we reject {reject/fail to reject} the null hypothesis."**
  - This is "drawing a relevant conclusion"

If I want to see if Harvard students get less sleep than other college students, what should my hypotheses be (in terms of pop. parameters)?

Question:

If I want to see if Harvard students get less sleep than other college students, what should my hypotheses be (in terms of pop. parameters)?

We have a **binary categorical explanatory variable** (Harvard or not) and **numerical response variable** (hours of sleep). This is a **one-tailed difference of means**.

$\mathbf{H_o}$: $\mu_{\text{Harvard}} - \mu_{\text{Other}} = 0$ (Harvard students get same amount of sleep)

$\mathbf{H_A}$: $\mu_{\text{Harvard}} - \mu_{\text{Other}} < 0$ (Harvard students get less sleep)

If I observe a difference of means of -2.7 hours (and a p-value of 0.003), what is an interpretation of the p-value and a conclusion? Assume $\alpha = 5\%$.

Question:

If I observe a difference of means of -2.7 hours (and a p-value of 0.003), what is an interpretation of the p-value and a conclusion? Assume $\alpha = 5\%$.

Using the p-value formula...

If **there was no difference in mean hours of sleep between Harvard and non-Harvard students**, then the probability of observing our **test statistic, a difference of -2.7 hours**, or **less** would be 0.3%.

Because 0.3% is a **low** probability (0.3% < 5%), we **reject** the null hypothesis.

qnorm() or qt()?

- Recall **means, difference in means, and linear regression** use the **t distribution**—thus, `qt()` is appropriate for finding critical values
  - If you don't remember, check this guide: https://drive.google.com/file/d/1O6gzE4jXfFCyKgBBPO17jfM8NxI9yMZy/view?usp=sharing

I want to calculate a confidence interval for how much more money Harvard students spend than non-Harvard students. Would I use qt() or qnorm()?

Question:

I want to calculate a confidence interval for how much more money Harvard students spend than non-Harvard students. Would I use qt() or qnorm()?

qt()!

Here, we're working with a **difference in means**, so the standardized test statistic follows the **t distribution**.

Questions?

Let’s get some practice coding!